Summary Report

Medical Effectiveness Research Data Methods

RA
399
A1
M435
1992

U.S. Department of Health and Human Services
Public Health Service
Agency for Health Care Policy and Research

Department of Health and Human Services
Louis W. Sullivan, M.D., *Secretary*

Public Health Service
James O. Mason, M.D., Dr. P.H., *Assistant Secretary for Health*

Agency for Health Care Policy and Research
J. Jarrett Clinton, M.D., M.P.H., *Administrator*

Office of Science and Data Development
J. Michael Fitzmaurice, Ph.D., *Director*

RA
399
.A1
M435
1992

Summary Report

Medical Effectiveness Research Data Methods

Editors:

Mary L. Grady
Harvey A. Schwartz, PhD, MBA, MS

U.S. Department of Health and Human Services
Public Health Service
Agency for Health Care Policy and Research
Rockville, MD

AHCPR Pub. No. 92-0056
July 1992

The opinions expressed herein are those of the authors. Publication of this report does not necessarily represent approval, concurrence, or official endorsement of the statements and opinions herein by the Agency for Health Care Policy and Research, the Public Health Service, or the U.S. Department of Health and Human Services.
This document is in the public domain and may be used and reprinted without special permission, except for those copyrighted materials noted for which further reproduction is prohibited without the specific permission of the copyright holders. AHCPR will appreciate citation as to source, and the suggested format is provided below.

Grady, M.L. (Ed.) *Medical Effectiveness Research Data Methods.* AHCPR Pub. No. 92-0056. Rockville, MD: Agency for Health Care Policy and Research. June 1992.

Acknowledgments. We gratefully acknowledge the assistance of Doris Biscarr, Office of Science and Data Development, Agency for Health Care Policy and Research, for her tireless efforts in coordinating the conference; Naomi Levy for editorial assistance; and Moshman Associates for logistical support.

Foreword

This volume features presentations from a conference on data methods for medical effectiveness research, sponsored by the Agency for Health Care Policy and Research in September 1991 in Minneapolis, MN.

These papers present practical, tested methods for working with large data bases and combining data from disparate sources to facilitate use of the data in medical effectiveness research. Problems in accessing and using such data, as well as suggestions for overcoming these problems, are discussed.

We trust that the guidance offered within these pages will prove useful, both to health services researchers and to those who design and maintain health care data bases.

J. Jarrett Clinton, M.D., M.P.H.
Administrator
Agency for Health Care Policy and Research

Contents

Overview

Medical Effectiveness Research Methods and Data 1
Harvey A. Schwartz, PhD, MBA, MS

Keynote Address

What Do We Do When the Data for Medical Effectiveness Research Are Not Perfect? 5
J. Michael Fitzmaurice, PhD

Issue

The Thrombolytic Predictive Instrument (TPI) Project: Combining Clinical Study Data Bases to Take Medical Effectiveness Research to the Streets 9
Harry P. Selker, MD, MSPH; John L. Griffith, MS; Joni R. Behansky, RN, MPH; Robert M. Califf, MD; Ralph B. Agostino, PhD; Michael M. Laks, MD; Kerry L. Lee, PhD; Charles Maynard, PhD; Galen S. Wagner, MD; W. Douglas Weaver, MD

Discussion

Combining Data from Different Studies 33
Mark A. Hlatky, MD

Issue

Research on Outcomes of Care Using a Medicaid AIDS Data Base: Overcoming the Curses of Imperfect Data 37
Barbara J. Turner, MD, MSEd; Leona E. Markson, ScD; Robert Houchens, PhD; Thomas Fanning, PhD

Discussion

Overcoming the Curses of Imperfect Data 53
Fred J. Hellinger, PhD

Issue

Getting the Most Out of Messy Data: Problems and Approaches for Dealing with Large Administrative Data Sets 57
Patrick Romano, MD, MPH; Harold S. Luft, PhD

Discussion

Using Physician Claims to Supplement Hospital Data 77
Janet B. Mitchell, PhD

Issue

Risk Adjustment for Medical Outcome Studies 83
Lisa I. Iezzoni, MD, MSc

Discussion

Risk Adjustment for Medical Outcome Studies: Are We Opening Pandora's Box? 99
Constantine A. Gatsonis, PhD

Issue

The Relationship Between Use of Health Care Services and Health Status: Dilemmas in Measuring Medical Outcome in Low-Income Racial/Ethnic Populations 103
Ruth E. Zambrana, PhD

Discussion

Dilemmas in Measuring Medical Outcome in Low–Income Racial/Ethnic Populations 115
Jose J. Escarce, MD, AM

Issue

Record Linkage: An Overview 119
Leslie L. Roos, PhD; Andre Wajda, MSc; J. Patrick Nicol, BComm; Jan Roberts, MD

Discussion

A Discussion of Record Linkage 131
Martha E. Fair, MS

Issue

Survival Analysis Methods for Medical Treatment Effectiveness Research 137
David C. Naftel, PhD

Discussion

Problems and Potentials in Modeling Survival 151
L. Richard Smith, PhD; Frank E. Harrell, Jr, PhD; Lawrence H. Muhlbaier, PhD

Issue

Episode of Care Framework: Utility for Medical Effectiveness Research 161
Donald M. Steinwachs, PhD

Discussion

Using Episodes of Care in Effectiveness Research 173
Emmett B. Keeler, PhD

Summaries

Medical Effectiveness Research Data Methods: Summary and Reactions 177
Joseph P. Newhouse, PhD

Learning from the Experts: What We Know About Effectiveness Research 183
Barbara Starfield, MD, MPH

Medical Effectiveness Research Methods and Data

Harvey A. Schwartz, Ph.D., M.B.A., M.S.

Introduction

There are many books on how to conduct research. There are also numerous books on the methodology of research designs, but they focus chiefly on data collection, not data analysis. Many books that do tackle analysis concentrate on mechanical procedures and not on the logic of interpretation. Equally, the even less extensive literature on interrelation of data derived from research methods is seldom explicit about analytic procedures.

Researchers can hope to overcome the intrinsic bias that comes from single-method, single-observer, single-theory studies by combining multiple observers, theories, methods, and data sources. Although this seems to provide a warrant for eclecticism, little help has been extended to those who try it, piling up reams of data with disparate epistemologic underpinnings but no guidance on how to reconcile these data. A researcher empowered by meta-analytic tools can integrate, synthesize, and make sense of collections of empirical findings from individual studies. Nevertheless, there is disagreement on exactly what to do should the collection of studies be significantly heterogeneous (Cook and Leviton, 1980; Hedges and Olkin, 1985).

Advocates of particular methodologies have been concerned more with asserting or defending their accustomed lines of inquiry than with indicating the possible points of convergence with other approaches. It is hard to see how a suggestion that researchers "approach data with multiple perspectives and hypotheses in mind" amounts to much more than an injunction to be imaginative in pondering what their data might mean (Hammersley and Atkinson, 1983). Such homilies pay too little attention to how researchers deal with discrepancies between discrete items of data derived from the same method, much less cases of intermethod discrepancy. Certainly, deviant cases can be very useful analytically, but the processes by which the researcher decides that this is the exception that refines (if not proves) the "rule"—rather than the case that implies a whole new rule altogether—are seldom made explicit.

Dr. Schwartz is a Senior Economist, Office of Science and Data Development, Agency for Health Care Policy and Research.

Data-Related Methods in Medical Effectiveness Research

Medical effectiveness research methods are the means for assessing and improving the development of data for medical effectiveness research. This entails examining scientifically how real-world observations are transformed into data sets that describe health care and the clinical and functional status of patients. It also involves developing methods for assessing and improving the efficiency and accuracy of this transformation.

Any empirical observation usually provides the observer with much information. Typically, however, only certain aspects of this information will be useful for medical effectiveness research. The medical effectiveness researcher takes a vitally important step in his or her analysis simply by defining the information to be captured and by culling out those pieces of information to be used from all those considered. The information that is used comprises the data, and it is clearly only a subset of observable reality and of information that could be captured. Hence, it is important to distinguish between observations (the information that health care providers and patients see in the real world around them) and medical effectiveness research-related data (information that the medical effectiveness researcher chooses to analyze).

A major concern of medical effectiveness research methodologists' scientific examinations is to specify how the latter are derived from the former and to determine and disseminate the implications for medical effectiveness research findings. This specification is clearly an important step towards improving the methods by which patient care research-related data can be derived. Ultimately, improving these methods should

lead to the delivery of high-quality data for conducting significant medical effectiveness research.

Future Developments

Methodologists need to explore new and more productive techniques for obtaining and analyzing health care data. More clinical data will probably be used in the conduct of medical effectiveness research; administrative data probably will be less important until their accuracy is determined to be high and they can be merged with clinical data. It is very unlikely that there will be data from randomized controlled trials to meet all of the avenues of research that the medical effectiveness program will want to explore. Accordingly, there will be an increasing reliance on observational data of a clinical nature. Since there will be increasing emphasis for analyzing such data and ascertaining its validity, more statistical research in this area is necessary.

Considerations of privacy and confidentiality may preclude the sharing of medical records containing patient identifiers (Office of Science and Data Development, 1991). Methodologies that improve accuracy when linking records based on data other than the patient's name or hospital record number may prove to be critical for achieving significant medical effectiveness research. Some examples of these methodologies are a computer system for generalized record linkage under conditions of uncertainty, under the name of UNIMATCH (for UNIversal MATCHer) (Jaro, 1972), and the EM algorithm (a matching algorithm which uses the linear sum assignment model to "pair" records) (Jaro, 1989). Successful research will allow longitudinal analysis of serial episodes of care with single institutions or among a number of institutions.

In patient outcomes studies, it is often necessary for the researcher to acquire data to follow the patient beyond the proximate outcome (outcome at discharge from the hospital or on exit from the physician's office or clinic). Instruments and techniques are required to permit the acquisition of valid followup data to ascertain the ultimate (or at least the longer term) effectiveness of the intervention. Development and testing efforts are required to arrive at the most efficient ways for acquiring this crucial data.

Until the time when patient care data are entered into automated information systems and are rapidly retrievable with uniform definitions and formats, there must be advances in the technology used for obtaining information from paper records. Further developments are needed in information science technology to permit extraction from existing medical records of data useful for medical effectiveness research. Some examples of this technology are the extraction algorithms developed with the initial support of the Health Care Financing Administration: the Uniform Clinical Data Set and the Ambulatory Record Project. Investigations of the feasibility of applying automated data extraction algorithms and procedures for clinical data elements useful for medical effectiveness research—including the accuracy and cost-effectiveness of data extraction techniques—are also needed.

Medical Effectiveness Research Data Methods Conference

Although medical effectiveness research as a field of study with goals and guidelines is new, the papers at this conference demonstrate the extraordinary efforts now contributing to the development of improved methodologies in medical effectiveness research. Regardless of the accomplishments in medical effectiveness research, the development of methodologies with respect to data sources as a field of study is new. The presenters and the papers emphasized (1) the word "using" in the conference theme—"Methodologies for using various data sources in medical effectiveness research"—and (2) improving or enhancing methodologies for using data sources in medical effectiveness research.

According to *Webster's II New Riverside University Dictionary* (Soukhanov and Ellis, 1984), use means to bring or put into service or action; to put to some purpose. The concept is different from consume—that is, to use up or to waste away. However, it is possible to consume data—that is, use up the data without attempting to address preconceived research questions or without testing a preconceived model.

Each paper in this volume addresses a well-defined topic. Many are concerned with methodologic issues pertaining to error, that is, deviation from truth, accuracy, or correctness. Medical effectiveness researchers could deal with methodologic issues related to error and other issues as they see fit and continue on with their studies. The quality of the papers, however, and the number of participants who chose to attend the conference clearly support the initiative to discuss error-related issues and associated methodologic problems in a public forum. I have no doubt that medical effectiveness research will benefit from our deliberations and papers through stimulation, exchange of relevant experience, and generation of important new ideas, which may pertain to analytic procedures with respect to the interrelation of data.

References

Cook, T. and Leviton, L. (1980). Reviewing the literature: A comparison of traditional methods with meta-analysis. *Journal of Personality*, 48, 449-472.

Hammersley, M. and Atkinson, P. (1983). *Ethnography: Principles in practice.* London: Tavistock.

Hedges, L. and Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Jaro, M.A. (1972). *UNIMATCH: A computer system for generalized record linkage under conditions of uncertainty* (Reprinted from AFIPS, Conference Proceedings, Vol. 40). Montvale, NJ: AFIPS Press.

Jaro, M.A. (1989). *Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida* (Reprinted from the *Journal of the American Statistical Association* (1989), 84(406), Applications and case studies). Alexandria, VA: American Statistical Association.

Office of Science and Data Development. (1991). *The feasibility of linking research-related data bases to Federal and non-Federal administrative data bases: Report to Congress.* Rockville, MD: Agency for Health Care Policy and Research.

Soukhanov, A.H. and Ellis, K. (Eds). (1984). *Webster's II New Riverside University Dictionary.* Boston: Houghton Mifflin Company.

What Do We Do When the Data for Medical Effectiveness Research Are Not Perfect?

J. Michael Fitzmaurice, Ph.D.

Introduction

As our Nation faces the challenge of attempting to leverage the enormous power of its health care system by improving its effectiveness, we are drawn together at this conference by the potential that better methods and better data will lead to better patient outcomes. There is a strong national interest in the quality of health care in this country. In the past 20 years, research on medical practice variations, the appropriateness of medical care, and patient outcomes has drawn widespread attention to the variation in medical practice patterns for specific illnesses across similar communities.

There is also a strong national interest in the cost of health care. When the cost of health care goods and services produced in the United States in 1990 are aggregated, they come to $671 billion; that is 11.1 percent higher than in 1989 and 12.2 percent of our gross national product (GNP). In the year 2000, national health expenditures are projected to be $1.616 trillion dollars and 16.4 percent of the GNP. It is clear that our society must obtain the maximum value for its dollars spent on health care.

In response to these and other health care concerns, the Agency for Health Care Policy and Research (AHCPR) was created by Congress in 1989:

> . . . to enhance the quality, appropriateness, and *effectiveness* of health care services and access to such services. (Omnibus Budget Reconciliation Act (OBRA), 1989)

Congress intended for us to accomplish this ". . . through the establishment of a broad base of scientific research and through the promotion of improvements in clinical practice and in the organization, financing, and delivery of health care services" (OBRA, 1989).

Dr. Fitzmaurice is Director, Office of Science and Data Development, Agency for Health Care Policy and Research.

The outcomes research mandate of AHCPR is to: (1) identify the most effective methods for preventing, diagnosing, treating, and managing various health conditions; (2) issue condition-specific practice guidelines; and (3) evaluate alternative health care services and procedures.

The basic underpinning of the condition-specific practice guidelines is medical effectiveness research (clinical trials and other studies). And the foundation of medical effectiveness research is good scientific method robustly applied to good, well-understood patient care data.

Vision of a Perfect Patient Care Research Data Base

Let me construct for you a brief vision of a perfect patient care research data base. First, you would pose the research question: What are the expected relative outcomes of alternative treatments for a person with a given medical condition and a given set of risk factors? That is, with what probabilities do these outcomes occur and how do their expected occurrences depend on the risk factors? The ideal data base would include:

- Longitudinal data about individuals (person-based data), their medical conditions, and the evidence leading to their diagnosis. These data would be linked for specific patients over time and would join together similar patients, with "similar" defined by the specifics of the researcher's question.
- Data describing their treatment(s) in sufficient detail for analysis.
- Data on these patients' outcomes (health status, functional status, quality of life) at different points in time.
- Data measuring all of the risk factors thought to be associated with the treatments and patient outcomes: patient characteristics (past medical history, sociodemographic characteristics), site factors (volume/

performance measures, accessibility to pertinent diagnostic, technology, and treatment opportunities), and other factors.

The data elements would be uniform in definition and in coded values.

The ideal data base would be free from measurement errors and errors of omission. It would be representative of all the people in the community or at least a large percentage of them. And, you would know the population to which your study applied by the way in which this data base was compiled and described.

Society's concern about the confidentiality of this patient care data base would be addressed by sophisticated access restrictions and by severe penalties for violating the conditions under which the data could be used.

Medical Effectiveness Research Models

Unfortunately, the ideal data base for studying effects of medical treatments in the community is rarely available. In the absence of being able to design and carry out a large number of clinical trials to test medical effectiveness research hypotheses, the medical effectiveness researcher must develop a theoretical framework that accounts for the relation between treatments and patient outcomes.

This is not all—the researcher must also be aware of how existing confounding factors affect treatment choices and patient outcomes. Without sufficient data to measure all the risk factors thought to be associated with the treatment choice and patient outcomes, there is danger of over- or underestimating the effect of a given treatment choice on patient outcomes. That is, biased estimates of the probabilities of favorable (or unfavorable) patient outcomes may result from ignoring the effects of factors that influence both treatment choice and patient outcomes.

If the data available were perfect, complete, and accessible, we would be conducting community clinical trials. What do you do when the data are not perfect? Not only do you try to get better data, you also develop techniques to uncover the data problems and to handle them with soundly applied methodology.

Needs and Challenges

A number of needs and challenges are before us; these include the need to: (1) understand the data, the purpose for which they were collected, how they were collected, how they were checked for errors and corrected, how the elements were coded, and what the codes were intended to mean; (2) determine how the data failings may influence research results; (3) develop methods to reduce bias and large random variation and to obtain better patient outcome measures; and (4) develop rules for producing better data.

The uses of medical effectiveness data are important for addressing the critical challenges that we face. These challenges include:

- Improving knowledge of the effects of alternative medical treatments in the community;
- Improving the content of information by which to make and evaluate the appropriate choice of medical care for the patient;
- Improving knowledge of the effects of patient risk factors and their interactions on patient outcomes; and
- Assessing a massive arsenal of medical technology with sufficient data and analysis to compare one technology with the others.

We must meet these challenges to make our health care system more effective in addressing society's problems of access, quality, and the cost of health care.

Because original research data bases are expensive to assemble, existing analytical methods should obtain as much information from them as possible. Also, because of this expense, the usefulness of existing data bases should be maximized, including sharing of data. Further, the systems that produce these data bases should be flexible enough to permit additional data collection for testing new hypotheses, as medical effectiveness research findings become widely known. Many patient care data bases are not collected with research as their primary purpose but could be improved continuously as medical effectiveness research findings demonstrate the importance of more accuracy in data recording and the need to record variables that are found to be significant influences on the probabilities of favorable outcomes.

In the Public Health Service report, *Healthy people 2000, national health promotion and disease prevention objectives* (1991), the Secretary of the Department of Health and Human Services, Louis W. Sullivan, M.D., states:

> . . . personal responsibility . . . truly is the key to good health. Evidence of this . . . abounds in our concern about the dangers of smoking and the abuse of alcohol and drugs; in the emphasis we are placing on physical and emotional fitness; in our growing interest in good nutritional practices; and in our concern about the quality of our environment. (page v)

To meet the challenges facing our Nation today, I expect patient outcome research designs will include more and more patient characteristics describing lifestyle influences on patient outcomes. Some of these influences are cigarette smoking, excessive drinking, drug abuse,

and poor nutrition. Socioeconomic characteristics may well prove to be important in medical effectiveness research, if collected with accuracy.

Additional Methodological Challenges

There are many additional methodological challenges facing medical effectiveness researchers and methodologists; these include: How can patient care data from different studies be pooled with good effect? Is pooling possible only for clinical trial data, or for community data as well? How can there be better development and application of methods for using pooled data and for determining bias in the resulting pooled data base?

How can there be more compatibility between administrative data produced about patient care and clinical data produced for medical effectiveness research? How can we improve the content, linkability, and accuracy of existing systems of patient care data? How can we improve the accessibility of these data for research purposes? How can better measures of patient outcomes be developed and collected?

Here and now, we must deal with existing data sets and the data sets we will build in the near future. The lessons learned here, however, will have strong implications, not only for the development and application of good methods to existing data, but also for the development of improved methods and better data in the future.

Conclusion

The presentations at the conference, and the papers in this volume, demonstrate a high caliber of research. One of the goals of the conference was to prepare the groundwork for better-informed clinicians and patients—those who have the necessary information for considering different treatment options for a condition and for reaching mutually effective decisions.

Significant progress was made during the conference toward understanding what the data problems are, which data methods are used to confront these problems, and where efforts may be productively directed in the future.

References

Healthy People 2000. National Health Promotion and Disease Prevention Objectives (DHHS Pub. No. 91-50212). (1991). Washington: Public Health Service.

Omnibus Budget Reconciliation Act of 1989. (1989). Public Law 101-239, Title IX, Public Health Service Act, Section 6103. Washington: U.S. Government Printing Office.

The Thrombolytic Predictive Instrument Project: Combining Clinical Study Data Bases to Take Medical Effectiveness Research to the Streets

Harry P. Selker, M.D., M.S.P.H.
John L. Griffith, M.S.
Joni R. Beshansky, R.N., M.P.H.
Robert M. Califf, M.D.
Ralph B. D'Agostino, Ph.D.
Michael M. Laks, M.D.
Kerry L. Lee, Ph.D.
Charles Maynard, Ph.D.
Galen S. Wagner, M.D.
W. Douglas Weaver, M.D.

Introduction and Background

Thrombolytic therapy (TT) may well represent the most effective single treatment devised for treating acute myocardial infarction (AMI), and it thus has enormous potential impact on this most common cause of death in the United States. Patients' outcomes of TT, however, depend entirely on getting the treatment to appropriate candidates within the first hours of AMI symptoms. If given beyond 12 hours after symptom onset or to patients with very small AMIs, there is little likelihood of benefit but substantial risk of complications and cost.

In order to maximize the effectiveness of TT in widespread clinical use, a method must be developed that:

1. Specifically identifies those patients most likely to benefit from TT in the real-time clinical setting, and
2. Facilitates the earliest possible administration of TT.

Additionally, a set of AMI outcome predictors—accurately adjusted for risk—is needed to objectively monitor TT's effectiveness and provide useful performance feedback.

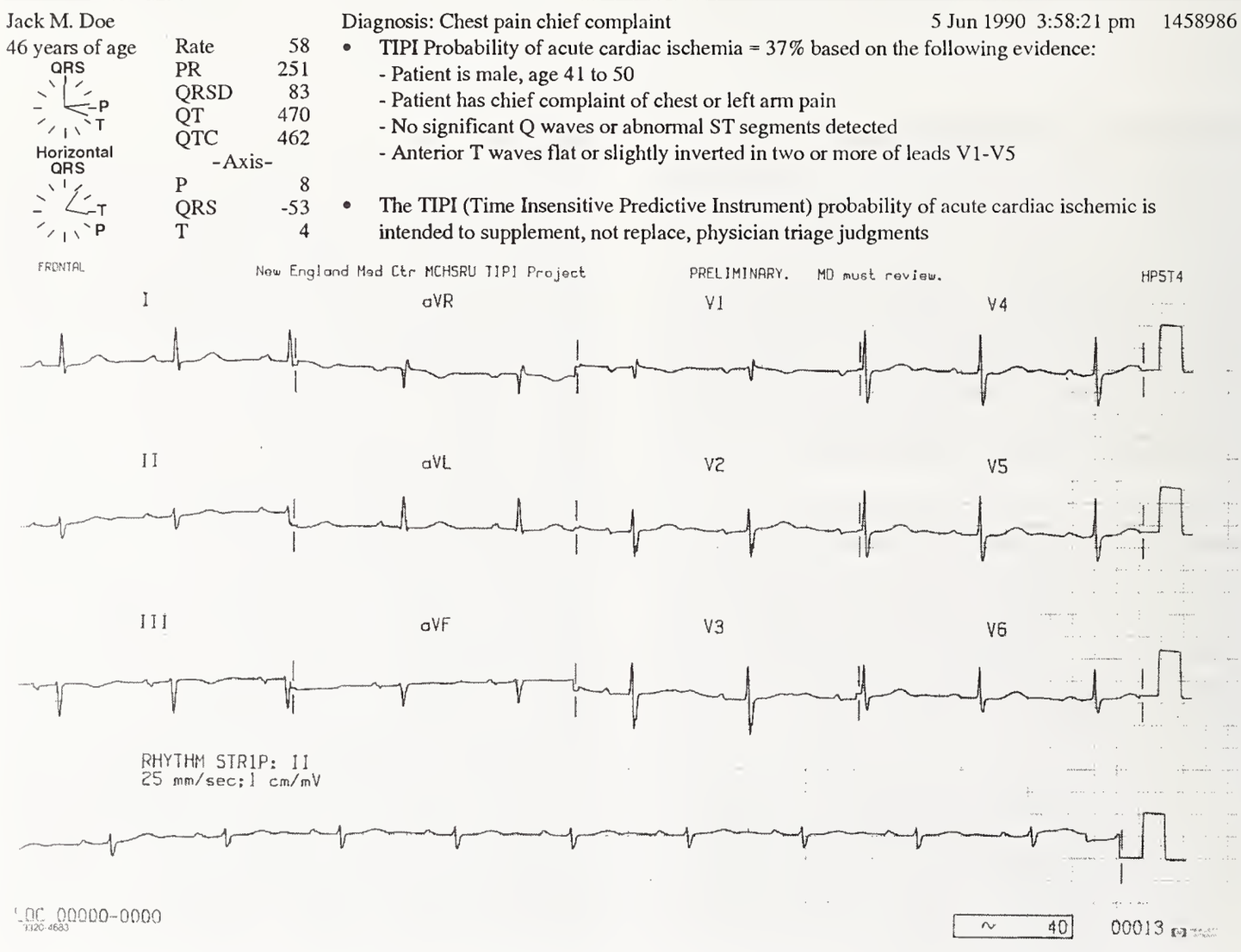
Over the past decade, we have developed "predictive instruments" for acute cardiac ischemia (acute cardiac ischemia here refers to AMI as well as new onset or unstable angina pectoris, together representing the group physicians generally consider for coronary care unit [CCU] admission) (Pozen, D'Agostino, Mitchell, and others, 1980; Pozen, D'Agostino, Selker, and others, 1984; Selker, D'Agostino, and Laks, 1988; Selker, Griffith, and D'Agostino, 1991). Based on logistic regressions, use of the instruments requires an indication of the presence or absence of seven clinical features in a given patient to calculate the 0-100 percent probability that the patient truly has acute ischemia. This probability then can be used in the emergency clinical setting to aid physician CCU admission decisionmaking. In practice, for example, if using the version incorporated into a computerized electrocardiograph, the user inputs into the electrocardiograph the patient's age, gender, and whether there is a chief complaint of chest pain; the electrocardiogram (ECG) variables are automatically acquired by the electrocardiograph. The probability of acute ischemia together with the patient's characteristics used for the calculation are then printed on the top of the ECG (Figure 1) (Selker, D'Agostino, and Laks, 1988; Selker, Griffith, and D'Agostino, 1991).

Dr. Selker is Director, Center for Cardiovascular Health Services Research and Center for Health Services Research and Study Design, New England Medical Center, Tufts University. Mr. Griffith, Ms. Beshansky, and Dr. D'Agostino are also with the Center. Dr. Califf, Dr. Lee, and Dr. Wagner are with the Department of Medicine, Duke University Medical Center. Dr. Maynard and Dr. Weaver are with the Department of Medicine, University of Washington School of Medicine. Dr. Laks is with the Department of Medicine, UCLA-Harbor Medical Center.

In a multicenter trial, including hospitals ranging from urban major teaching centers to rural nonteaching hospitals, the use of the acute ischemia predictive instrument reduced by 30 percent the unnecessary CCU admissions of patients without acute ischemia but did not increase the number of patients with acute ischemia not admitted to the CCU. If the predictive instrument were used widely in this country, such use would translate into 250,000 fewer unnecessary CCU admissions each year, corresponding to a savings of approximately $1 billion in hospital charges. Thus, the use of such a clinical outcome predictor in real time may provide an opportunity to improve the appropriateness of care and avoid unnecessary costs, and it can be considered as an applied form of effectiveness research (Institute of Medicine, 1990).

Of note, a newer version of this decision aid is a "time-insensitive" predictive instrument (TIPI). It uses only clinical variables that can be acquired accurately both in the first 10 minutes of routine clinical care and also retrospectively by medical record review. Thus, in addition to real-time clinical advice, the TIPI can also provide accurately risk-adjusted clinical outcome predictions that can be used to judge emergency physician and hospital performance retrospectively (Selker, Griffith, and D'Agostino, 1991). An analogous TIPI for AMI acute mortality has been developed and used to compare hospital performance (Selker, Griffith, and D'Agostino, 1991b). In addition to the clinical utility of the real-time probability, the availability of a clinically valid risk-adjusted outcome measure, based only on the

Figure 1. Computerized electrocardiograph version of Time-Insensitive Predictive Instrument (TIPI) for acute cardiac ischemia



first 10 minutes of care, greatly enhances the possibility of cooperation among clinicians, hospitals, payers, and monitoring agencies (Selker, Griffith, D'Agostino, 1991b).

Analogous to our earlier predictive instruments for acute cardiac care, a decision aid that could provide emergency department (ED) and emergency medical service (EMS) clinicians with predictions of the likely outcomes of using TT, as well as risk-adjusted outcomes for judging care retrospectively, would be of considerable utility. Specifically, such a thrombolytic predictive instrument (TPI) would be most helpful if it provided the clinician with predictions of the likely impact on mortality (acute and long-term) and the likelihood of serious complications (for example, major bleeding, stroke), conditioned on the specific circumstances of the patient about to be treated.

Herein we describe our approach to creating such an instrument, based on multivariate regression formulas that use variables readily available within minutes of a patient's presentation. The TPI according to our scheme should (1) specifically identify those patients most likely to benefit from TT and (2) facilitate the earliest possible administration of TT. Additionally, when the data are saved in a data base, the TPI will provide risk-adjusted outcomes of the use of TT for retrospective assessment of care effectiveness. Six major advantages that should optimize outcomes of TT are thereby sought:

1. Because the TPI will be directly modeled on the actual data from some of the Nation's major just-completed and ongoing TT trials and from studies of analogous patients in the immediately pre-TT era, we will avoid the enormous costs and substantial time of collecting such data and be able to concentrate on their optimal use to make the most generalizable and clinically useful tool possible.

2. Because the TPI will identify candidates for TT based on a multivariate integration of a number of attributes and will assess these attributes in their most useful format (for example, the actual ECG wave form available to the computerized electrocardiograph rather than a physician's use of only several standard ECG-based criteria for TT use), the identification of potential candidates for TT will be refined for real-time use. Thus, the TPI will likely identify TT candidates not identified by current criteria (that is, increase sensitivity), which should expand the pool of patients benefitting from TT.

3. Again, because of its integration of multiple factors influencing benefits and complications of TT, the TPI will also better identify patients unlikely to benefit from TT, including those likely to suffer complications. Thus, it will likely identify patients who should not receive TT but would have met conventional treatment criteria (that is, increase specificity). Thus, the TPI should reduce adverse effects and costs from unnecessary TT use.

4. Because the TPI will identify TT candidates at the earliest possible time when it is incorporated into a computerized electrocardiograph for EMS and ED use, the TPI should expedite early administration—a major determinant of the effectiveness of TT—and thereby improve TT's efficacy.

5. Based on the data from the Thrombolysis and Angioplasty in Myocardial Infarction (TAMI) Trial, which constitutes one of this study's major data bases, the TPI will be able to provide an additional decision-making aid for the use of acute supplemental angioplasty in addition to TT. (Because this represents treatment that is still not conventional, the decision to include this special component will be made by the study's senior clinicians and other experts at the time of the TPI's assembly.)

6. The creation and validation of the TPI's component outcome predictive instruments for the major outcomes of AMI and its therapy will be useful as tools for comparing outcomes of care among different providers and institutions. Of importance, because the predictive models will be of proven clinical validity, their use for judging care should facilitate communication among those who need to cooperate in improving care, namely clinicians, administrators, payers, and agencies (Selker, Griffith, and D'Agostino, 1991b).

Background on the Determinants of Outcome of TT for AMI: Possible Avenues of Improvement

Even in a field as replete with technological advances as the treatment of coronary heart disease, TT may offer the greatest potential for saving lives of any single modality yet devised. Controlled clinical trials have now established that if given very early in the course of certain types of AMIs, TT's impact on acute mortality may approach, or even exceed, a 50-percent reduction (Gruppo Italiano per lo Studio Della Streptochinasi nell'Infarcto Miocardico [GISSI], 1986; ISAM Study Group, 1986; ISIS Steering Committee, 1987; Simoons, Serruys, Van Den Brand, and others, 1986; Van de Werf, Ludbrook, Bergmann, and others, 1984; White, Norris, Brown, and others, 1987). Moreover, benefits also have been documented in increased patency of infarct-related coronary arteries (Bergman, Lerch, Fox, and others,

1982; Chesebro, Knatterud, Roberts, and others, 1987); improved cardiac function as indicated by left ventricular ejection fraction (LVEF) (Simoons, Serruys, Van Den Brand, and others, 1986); and, as would be expected in conjunction with greater LVEF (Norris, Barnaby, Brandt, and others, 1984), improved long-term mortality (Franzosi, Mauri, Pampallona, and others, 1987; GISSI, 1987; Simoons, Serruys, Van Den Brand, and others, 1986).

If patient outcomes in the general clinical use of TT reach this level of performance, the national impact will be substantial. Even if EMS does not significantly improve capture of the current large number of AMIs that never benefit from acute hospital care, more than 250,000 AMI patients are hospitalized in this country each year (American Heart Association [AHA], 1976), and a drop in acute mortality—from the typical 15 percent to 7.5 percent for appropriate candidates—would save thousands of lives annually, and additional patients would benefit from improved long-term survival.

TT use, however, should not be unrestrained. The complications can be serious; these include stroke, hemorrhage, and other significant problems. When TT is used on a widespread basis, these events will not be inconsequential. Additionally, given that TT medication and associated care are expensive—for example, tissue-type plasminogen activator (tPA) costs $2,200 for a single administration—its use leads to yet more costly intensive care hospitalization and either immediate or followup invasive tests and/or treatment. At New England Medical Center, it is estimated that the use of TT and its consequences increase AMI hospitalization charges by approximately $5,000 and actual hospital costs by about $4,000. Thus, the financial impact of widespread use of TT for acute care alone, without accounting for related additional coronary bypass surgery, would be on the order of over $1 billion yearly.

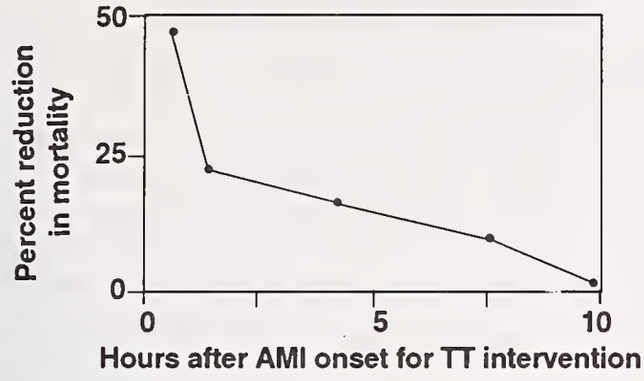
To balance these benefits, risks, and costs practically, in the midst of the current rapid proliferation of this new technology, a sensitive and specific method that can be used in the emergency clinical setting where these decisions must be made is greatly needed for selecting appropriate TT candidates based on likely outcomes. To consider how this might in fact be accomplished, one must consider the major determinants of TT outcome.

The clearest finding to come from recent work with TT is the critical importance of duration of time since AMI symptom onset until treatment as the overriding determinant of efficacy. Given that experimentally, infarction appears to be complete by 4 hours after occlusion (Marder and Sherry, 1988), this is not surprising. Early TT studies did not show the successes that have been seen in this decade probably because of their longer delays from symptom (and presumably ischemia and infarction) onset to treatment. The recent dramatic results of TT trials have been clearly linked to rapidity of TT administration, generally within 4 hours. Moreover, even within this period, there appears to be a further direct relationship between mortality risk and time to treatment. For example, Figure 2 shows that in the GISSI Trial mortality was reduced by 47 percent (15.4 percent controls, 8.2 percent SK, P=0.0001) in those treated with streptokinase (SK) within 1 hour, in those treated within 3 hours, by 23 percent (14.1 percent controls, 11.7 percent SK, P=0.0005), and in those treated between 3 and 6 hours, by 17.0 percent (14.1 percent controls, 12.6 percent SK, P=0.03) (GISSI, 1986).[1] These and similar results have convinced many that the primary advance to be made in the delivery of care for acute coronary disease is the development of methods by which TT can be administered at the earliest possible time in an AMI's course. For although clearly beneficial if given early, if given after as few as 6 hours after symptom onset, there is much less likelihood of benefit, but substantial risk of complications and increased costs remains.

A number of approaches have been used to obviate delays in TT administration. Just the use of patient transport by an EMS to the ED has been shown to improve the timeliness of TT administration. For example, in a study done at three community hospitals, time from AMI symptom onset to TT treatment was 1.7 hours for EMS-transported patients versus 3.3 hours for self-transported patients (Hartmann, McKeever, Bufalino, and others, 1986). Similarly, other investigators have shown that tPA can be safely administered by helicopter-based EMS, permitting administration of TT even earlier in the course of an AMI, with associated clinical benefit (Collen, Topol, Tifenbrunn, and others, 1984). They also demonstrated that time was more important than the involvement of a high-technology tertiary care team, by showing earlier tPA administration (2.1 hours after AMI symptom onset versus 3.8 hours, P<0.001) and greater coronary patency and LVEF when TT was given by a community hospital staff, as opposed to either awaiting the assistance of a helicopter-based team for TT administration or awaiting helicopter transport to a tertiaryhospital where TT was given (Topol, Fung, Kline, and others, 1986).

[1]This is consistent with results from the Thrombolysis in Myocardial Infarction [TIMI] Trial [Chesebro, Knatterud, Roberts, and others, 1987] for TPA administration, showing infarct vessel canalization of 81 percent for those treated within 2 hours of symptom onset versus 65 percent for those treated within 3.5 hours [P=0.057].

Figure 2. Percent reduction in mortality in GISSI trial



The appropriateness of TT use at community-based sites is further reinforced by results of the TAMI Trial showing that the general use of acute invasive interventions in conjunction with TT is not obviously better than TT alone (Topol, Bates, Walton, and others, 1987). As the TAMI investigators point out, given that 81 percent (4,639) of U.S. hospitals have no cardiac catheterization laboratory, and a third of the 19 percent (1,077) with laboratories do not do percutaneous transluminal coronary angioplasty (PTCA), the clear avenue for improved AMI outcomes for the general population is via TT at community sites. The success of community-based TT administration to 11,712 patients in 80 percent of Italy's CCUs in the GISSI Trial emphasizes the feasibility of community hospital-based TT administration (American Heart Association, 1976; Grim, Feldman, Martin, and others, 1987). All of the over 5,000 acute-care hospitals in the United States with CCUs should be able to administer TT. Taken together, these results suggest that the primary avenue for improving TT delivery should be via community hospitals and their local EMS systems.

Although TT administration at community hospital sites appears preferable to transport to tertiary care hospitals, even transport to a hospital itself may represent a critical delay in TT. Thus, the question is, can TT be safely administered in the field? This approach has been tried via physician-staffed EMS (Applebaum, Weiss, Koren, and others, 1986) and by having physicians go to patients' homes (Villemant, Barriot, Riou, and others, 1987). In both cases, there were substantial reductions in times to treatment. Investigators in Israel were able to reach patients suffering AMIs within 5 minutes of being called (range 1-10 minutes) and administered TT within 32 minutes of arriving at the patient's site (range 22-46 minutes) (Applebaum, Weiss, Koren, and others, 1986).[2] On a widespread basis, however, the transport of physicians to the sites of all AMIs is impractical. The question then becomes, can physician-level expertise be transported to the field in a more efficient way?

Capitalizing on the trend over the past decade towards telemetry of three-lead ECG data for EMS administration of advanced life support supervised by physicians via radio, Grim, Feldman, Martin, and others (1987) suggested that a 12-lead portable computerized electrocardiograph with cellular telephone transmission could allow remote site physician-guided TT decisions. In the Seattle-based Myocardial Infarction Triage and Intervention (MITI) Trial, this approach is just now being tested (Weaver, Eisenberg, Martin, and others, 1990). Indeed, the participation of the MITI investigators and the MITI project's data are central to the TPI Project's activities. To further improve outcomes of TT, we will have to capitalize and improve on this EMS transport-based ECG-guided approach to TT use. The TPI is envisioned for exactly this application.

If the first observation about TT's effectiveness has been its time-sensitivity, then the second major finding of clinical trials has been that the patients most likely to benefit from TT are those with thrombosis of a major coronary artery likely to cause a large (for example, anterior) AMI. This is not surprising for two reasons:

1. The higher frequencies of deaths associated with larger AMIs provide greater statistical power for detecting beneficial effects of TT; and

2. The greater amount of myocardium at risk provides a greater opportunity for immediate benefit by reopening a more important closed coronary vessel.

Thus in most trials, anterior AMIs have shown the greatest benefit from TT; significant benefit is often found for inferior AMIs with reciprocal anterior ST (S-wave/T-wave) segment changes, but little benefit has been noted for the smallest, limited inferior AMIs (GISSI, 1986; Kennedy, Martin, Davis, and others, 1988; Kennedy, Ritchie, Davis, and Fritz, 1983; Marder and Sherry, 1988). However, not all types of higher risk AMI patients benefit more; for example, in the GISSI study, anterior and multiple-location MIs did show benefit in reduced mortality with TT, but so did patients younger than 65, those with a first AMI, and those with Killip Classes I and II AMIs, all lower risk characteristics. However, females older than 65, with a second AMI and Killip Classes III and IV did not show TT benefit (GISSI, 1986; Simoons, Serruys, Van Den Brand, and others, 1986). Indeed, different studies have

[2]Patients waited an average of 33 minutes to call, suggesting another potential limiting factor that is currently being addressed, particularly by Seattle-based investigators.

yielded somewhat varying results on infarct types most likely to benefit beyond the anterior/multiple-sites/inferior gradient. It is clear that more information and better ways of integrating available prognostic information are needed to optimize selection of TT candidates.

Vermeer, Simoons, Bar, and others (1986) did use a multivariate approach to analyzing factors influencing intracoronary TT efficacy, creating three linear regressions to predict three outcome variables: serum alpha hydroxybutyrate dehydrogenase, LVEF, and 3-month mortality. The important variables in predicting all of these outcomes were: sum of ST-segment deviation on the first ECG (ST elevation for anterior MIs and inferior ST elevation plus anterior ST depression for inferior MIs), delay from symptom onset to admission, Killip Class III or IV, the use of TT, and interaction terms between TT and each of the other variables. However, neither measures of performance of the regressions nor validation of their predictions with observed results were presented. Moreover, no attempt was made to make an analogous model predicting complications or to determine threshold for predicted TT impact that would be clinically important. Nonetheless, this study is important to our plans, as it represents an attempt to create a multivariate view of the benefits of TT. Ideally this multivariate approach, with more complete outcome predictions, should be available in real time at the clinical site: the explicit goal of the TPI.

The third major issue addressed by TT trials is the occurrence of complications and adverse effects. Because of the potential seriousness of TT complications, patients at very high risk are generally excluded. Typical are criteria of the Western Washington Intracoronary SK Trial. Exclusion criteria were: (1) age greater than 75 years; (2) presence of severe disease in other organ systems likely to reduce longevity or the ability to cooperate or give informed consent; (3) prior SK therapy; (4) any contraindication to anticoagulation, such as active ulcer disease, diabetic retinopathy, or hypertension with diastolic pressure higher than 120mm Hg; (5) recent surgical procedure (2 weeks) or neurosurgical procedure or stroke (4 weeks); (6) trauma likely to result in bleeding; (7) presence of central venous lines placed in incompressible entry sites; and (8) prior coronary bypass surgery (Kennedy, Ritchie, Davis, and Fritz, 1983). This is clearly appropriate, although it obviously limits specific information about complications in such individuals; however, complications do occur even in seemingly appropriate candidates, and they remain a limiting feature in the use of TT.

Fortunately, TT-related strokes are uncommon, and the risk can be assessed from larger trials. Strokes attributable to intravenous (IV) SK occurred in 0.01 percent in the GISSI Trial (GISSI, 1986; Simoons, Serruys, and Van Den Brand, 1986) in which age was not an exclusionary factor, and cerebral hemorrhage occurred in 0.05 percent in the Intravenous Streptokinase in Acute Myocardial Infarction (ISAM) Trial (ISAM, 1986). In the Western Washington Intravenous SK Trial, which will constitute part of the TPI data base, complications occurred in 23 percent of the treatment group and 14 percent of the control group (p=0.03) (Kennedy, Martin, Davis, and others, 1988). Complications included bleeding in 3 percent of those receiving TT (versus 0.6 percent of the controls, p=0.07), allergic reactions in 2 percent (versus 0 percent, p=0.07) (Kennedy, Martin, Davis, and others, 1988), ventricular fibrillation in 11 percent (versus 4.5 percent, p=0.02), and one nonfatal hemorrhagic stroke. Typical of most recent experience, in the New Zealand trial of IV SK (White, Norris, Brown, and others, 1987), no deaths or cerebrovascular bleeds occurred in their 155 treated patients; but 2 percent had allergic reactions, 3 percent had minor bleeds, and 7 percent had hypotension (systolic pressure less than 80mm Hg)—a rate similar to that in the tPA European trial (Verstraete, Bernard, Bory, and others, 1985).[3] Similar results were found in the more recent GISSI II Trial (GISSI, 1990) and ISIS III Trial (Chesebro, Knatterud, Roberts, and others, 1987; Grines, Nissen, Booth, and others, 1990), with the exception of an increased incidence of hypotension and allergic reactions. Clearly, in order to optimize outcomes of TT—to the extent that complications can be identified as being more likely based on certain presenting clinical features—there would be great value in integrating the predicted likelihoods of the complications into a real-time decision aid to assist in decisions to use TT. This tool is an explicit goal of the TPI.

A still-emerging issue in the use of TT is its use in conjunction with other invasive procedures, particularly PTCA. Theoretically, those patients who have large infarct-related coronary vessels that do not open with TT would greatly benefit from prompt mechanical opening of the artery by PTCA. As discussed above, however, earlier TAMI results did not show an advantage of PTCA in association with helicopter transport for tertiary care, at least in part because of associated delays in TT administration. Nonetheless, based on the pathophysiology of acute infarction, many cardiologists already feel that for selected patients, PTCA in addition to TT is probably more effective than TT alone. In fact, data from the TAMI-2 (Topol, Fung, Kline, and others,

[3]This comparability between SK and tPA was also seen in the TIMI Trial, where occurence of adverse effects also did not differ between those given SK and tPA and was not higher in the elderly (Chesebro, Knatterud, Roberts, and others, 1987).

1986), as well as TAMI-5 (Topol, Califf, George, and others, 1988), support this view. Indeed, it now seems likely that the earlier work did not use this approach sufficiently selectively, and thus the benefits accrued to some individuals were washed out by the lack of benefit to others. Because the lack of opening in such a vessel is generally apparent from a patient's immediate clinical course following TT administration (Kircher, Topol, O'Neill, and Pitt, 1987) (for example, lack of prompt amelioration of chest pain) and ECG (for example, persistence of anterior ST elevation in a patient with a large evolving anterior AMI), these factors could be incorporated into multivariate models predicting mortality and complications which include interaction terms for the use of PTCA in addition to TT. The fact that the TAMI data (including the TAMI-5 data) constitute a major part of our data base on which the TPI will be based allows us to do exactly this in constructing the TPI. Our final decision to implement this additional PTCA-beyond-TT component will depend on expert opinion and the results of studies available at the time that the TPI is being assembled.

Based on the issues reviewed above, the TPI is envisioned as a way to directly address the major factors that limit outcomes of TT. Specifically, it will:

1. Identify those patients likely to benefit from TT, based on their predicted outcomes, including benefits and complications; and
2. Facilitate the earliest possible administration of TT.

The TPI's first attribute will result from its multivariate weighing of the important clinical factors influencing key outcomes (for example, time since symptom onset, age, history of prior infarction) and the provision of the probabilities of these outcomes to aid treatment decisions. This should improve patient selection for treatment. Current clinical criteria, now generally based on relatively simple ECG ST segment elevation in specified leads, probably miss individuals that might in fact benefit from therapy. Indeed, more clinical trials should be run to better describe those patients likely to benefit from TT; however, trials are expensive, take a long time to produce results, and will become increasingly more difficult to perform as TT comes into widespread use. Moreover, the general evidence for benefit is now so great that trials of TT versus conservative therapy are no longer feasible, except for small groups where the benefit remains unclear. In the meanwhile, the multivariate integration of known important clinical factors by the TPI could maximize our ability to discern potential TT beneficiaries from existing data and thereby improve outcomes of TT use.

The TPI's second major contribution is related to its extension of the ability to administer TT to the out-of-hospital patient site. When integrated into a computerized self-interpreting electrocardiograph to guide EMS personnel's TT decision (at least initially with radio-based physician supervision), it should enable the earliest possible administration of treatment and thus efficacy of TT.

Finally, the creation and validation of predictive instruments for the major outcomes of AMI and its therapy will be useful as tools for comparing outcomes of care by different providers and institutions. There is currently a great need for validated clinically based predictors of important outcomes of care for this most common cause of death (Selker, Griffith, and D'Agostino, 1991). Indeed, such tools will be more credible and more likely to be accepted by clinicians if they are also shown to have the performance necessary for real-time clinical use. The TPI's component outcome predictive instruments should represent just such a set of tools for AMI.

Combining Clinical Study Data Bases to Create the TPI Data Base: The Basis for Detailed Models of AMI and TT Outcomes

As indicated, the TPI will consist of a set of predictive models, developed primarily using multivariate logistic regression, providing an ECG-based estimate of infarct size as a measure of AMI severity and determining probabilities of:

1. Acute (30-day) AMI mortality;
2. Long-term (1-year) AMI mortality;
3. Cardiac arrest during present AMI;
4. Serious bleed from TT; and
5. Hemorrhagic stroke from TT.

All but the infarct size regressions will be used to compute the likelihood of the given outcome—both assuming the patient is given TT and assuming the patient is not given TT—to assist in decisions about whether TT should be used. The LVEF prediction is intended as an overall indication of a patient's AMI severity at the time of presentation, aside from the issue of TT, as a point of reference for the clinician. A mockup of how these outcome predictions will be presented to the clinician is shown in Figure 3.

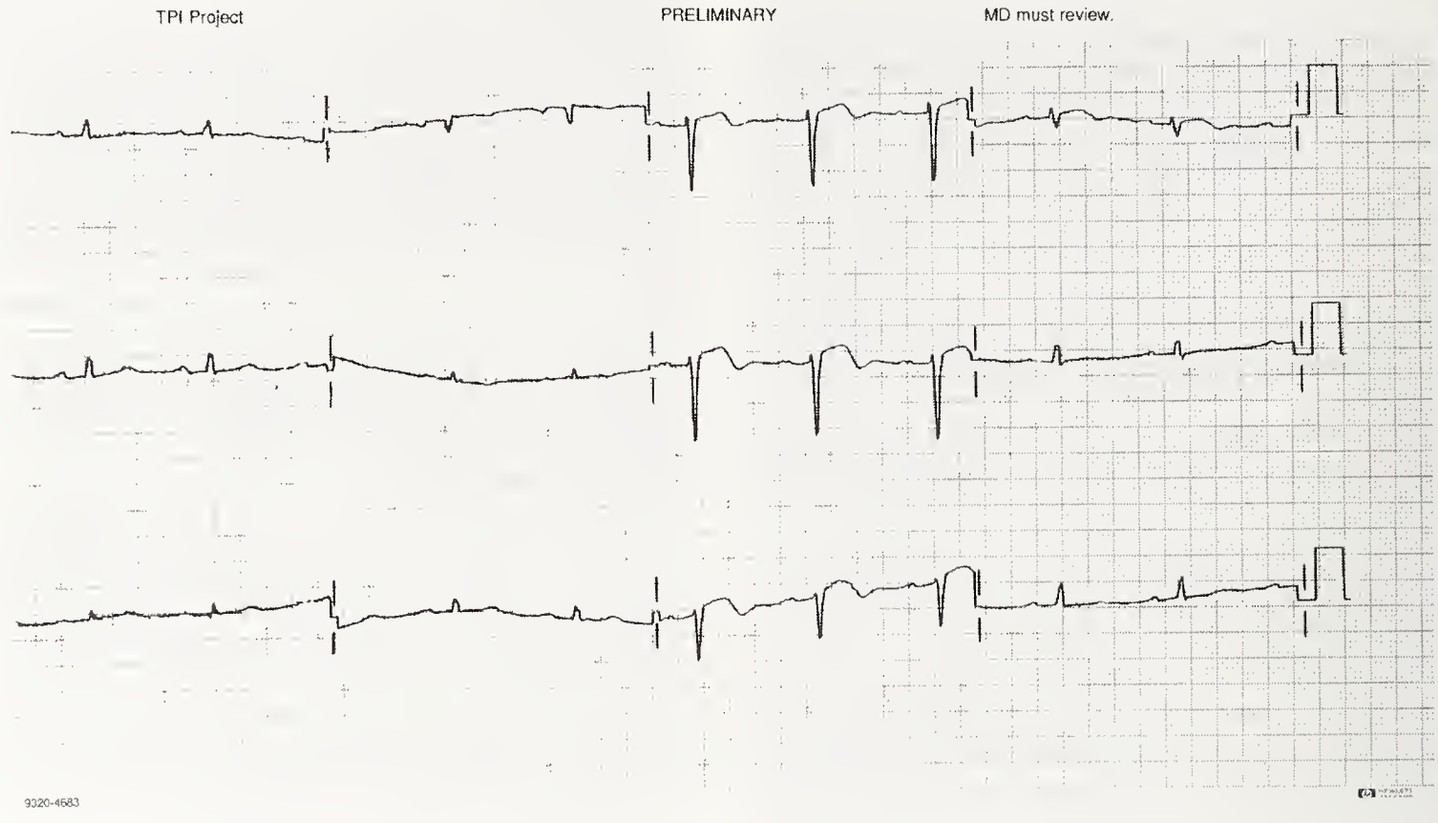
The TPI project, which will include creation of the component predictive models, their testing, their combination to create the TPI, and implementation of TPI in a computerized electrocardiograph as a clinical tool, is being conducted in five phases:

Figure 3. Mock-up of TPI electrocardiogram

John Doe 5 May 1991 1:43:23 pm

Thrombolytic Predictive Instrument (TPI):
(45 year old male, 3 hours from symptom onset, acute anterior MI, with ST elevation in leads V1, V2, V3, V4, no contraindications to thrombolysis, predicted post-MI LVEF= 40%)

Probabilities for thrombolytic therapy outcomes:
Acute (30-day) mortality: 6% if treated with thrombolysis, 12% if not.
1-year mortality: 8% if treated with thrombolysis, 17% if not.
Complications: Major bleed 3%; Hemorrhagic stroke 0.5%.



Phase One: Creation of the combined TPI data base from data bases from selected TT and AMI clinical trials and registries.

Phase Two: Development of the component AMI outcome predictive instruments that will constitute the heart of the TPI by using multivariate regression.

Phase Three: Testing and analyzing the predictive performance of the TPI component instruments on independent data sets.

Phase Four: Development and pilot testing of clinically useful forms of the TPI, including programming into a computerized self-interpreting electrocardiograph.

Phase Five: A multicenter prospective clinical trial of the impact of the TPI on TT use.

This paper focuses on Phase One and details our approach to combining clinical data bases as a foundation to create comprehensive models of the clinical outcomes of AMI and of TT use. Although the steps and examples detailed below relate to creating the TPI data base, the general approach should have utility in formulating the detailed models of medical treatment outcomes that are central to medical effectiveness research.

Advantages of combining data sets from different studies. A major advantage of combining data sets is the greater statistical power available for detecting and predicting differences in outcomes between those patients given and not given TT. This can be illustrated by considering the difference in power in modeling on one

of the Western Washington TT trials versus combining the two Western Washington TT trials.

Taking as an example a predictive model for acute AMI mortality with eight dichotomous independent variables, the statistical power for detecting a change in acute mortality from 20 percent to 15 percent increases from under 55 percent using one of these data bases to over 90 percent when they are combined. In a similar way, the power for finding a 5-percent change in the complication rate (from 17.5 percent in patients with TT to 12.5 percent in patients without TT, as was seen in the Western Washington IV SK Trial) increases from 60 percent to over 90 percent. In the case of the entire combined TPI data base, with over 4,000 subjects—as opposed to the 618 subjects yielded by combining the two Western Washington trials—the statistical power increases substantially. The size of the TPI data base, including the substantial number of observations in important subsets, provides the ability to carefully analyze variables, and interactions between variables, that otherwise would not be possible. This is particularly important for the TPI project, where we will be looking at the projected effectiveness of TT based on very specific clinical characteristics of patients.

Another major advantage accrued by combining these clinical data bases is the increased generalizability and robustness of the derived models. As in any multicenter data base, the development of a model using data acquired at different locations lessens the dependency on any particular data base and thereby reduces susceptibility to the idiosyncracies of specific data sets. Provided that the component data sets are compatible, this allows generation of a more generalizable and stable model.

A third advantage in this technique is the ability to develop models and test hypotheses that otherwise would be impossible. Because of the demonstrated efficacy of the use of TT and its widespread adoption in practice, it is no longer appropriate or possible to run a prospective trial in which patients are randomized to either receive or not receive TT. Thus at this point in time, the measurement of the benefits of TT for particular patient types based on comparisons of treated and untreated patients is only possible using data collected prior to the general use of TT.

Potential problems associated with combining data from separate clinical studies. Analogous to the situation of using data from different sites in a multicenter trial, the potential gain in statistical power and generalizability of a model constructed on many clinical trial data bases can only be achieved if the compatibility of the data sets is assured. A model constructed on a data base composed of noncompatible data sets or noncomparable variables may miscalculate the relative risk for an outcome related to a given variable (as expressed by the estimate of the beta coefficient in a logistic regression) or may erroneously determine the significance of a predictive variable (by miscalculation of a coefficient's estimated standard error). For the TPI—in which we will be comparing the predicted outcome for those receiving and not receiving TT—any such errors could be very serious.

In combining data sets, noncompatibility of data generally can occur because of:

1. Differences in outcome (dependent) variable definitions and in their collection;
2. Differences in independent variable definitions and methods of their collection; and
3. Differences among study protocols and populations.

Even for variables that seem unlikely to be problematic, such as "acute mortality" or blood pressure, how data were collected and exactly how a variable was defined can differ in critical ways among clinical studies. If variables are not redefined in a compatible way, the predicted relative risk may be erroneous, and statistical significance may be miscalculated due to the existence of spurious variability in the data. If not properly addressed, differences in study protocols—especially in explicit and implicit study inclusion/exclusion criteria, as well as differences in procedures and possible additional medical interventions—will add confounding factors that can bias the resulting models.

An additional concern to this project is the possible change in risk factors over historical time. Most of the available data on patients meeting eligibility criteria for TT who did not receive TT come from clinical trials done in the "prethrombolytic" era.[4] Thus, if changes in risk factors have occurred over time, this may confound the predicted differences in outcome between patients who received TT and those who did not.[5]

Our methods used in the creation of the TPI data base are designed to exploit the advantages of combining different studies' data while avoiding these possible pitfalls. This work, which constitutes Phase One of the TPI project, is outlined below. (The work of the other phases is subsequently described only very briefly to provide a context for our expectations of Phase One.)

[4]Such as the Center for Cardiovascular Health Services Research [CCHSR] data base and the earlier portions of the Duke CCU data bank.

[5]Included in the TPI data base are data from the Western Washington intracoronary and IV trials—in which contemporaneous TT treatment and non-TT control patients are present—which will assist in checking treatment effects in TPI models.

Table. Steps used to create the combined TPI data base

1.	Selection of participating clinical trials and data bases
2.	Development of a collaborative investigative team
3.	Preliminary determination of the availability and characteristics of data in each data base
4.	Specification of the dependent variables (outcomes)
5.	Determination and specification of the independent variables
6.	Specification of inclusive/exclusive criteria for treatment and control groups
7.	Addition of new data sets to the TPI Project
8.	Acquisition of new primary data
9.	Checking for consistency of effects across the TPI data base
10.	Division of the data base into development and test data sets

Phase One: Creation of the Combined TPI Data Base

The careful combining of data from the clinical trials and registry data bases may be viewed as involving 10 steps, as listed in the table. Steps are listed in rough sequential order, but often several steps are underway simultaneously. One step's activities may affect another, and the activities from different steps may be combined. Nonetheless, each step has an important role in creating the combined data base.

Step 1. Selection of participating clinical trials and data bases. From inception, it was clear that in order to develop the detailed outcome models for the TPI, we would need data from the major TT trials and registries in this country, along with other major studies of early identification and treatment of AMI. The selection of a data base for the TPI project was determined by the type of data collected, the patient population enrolled, availability of the data, and the interest and skills the study's investigators would bring to the project. We only considered prospectively collected data and only data sets for which the ECGs—the primary determinant of the use of TT—were available for checking and recoding. The use of data from such seminal studies provides the TPI project with well-documented and widely known data that have been reviewed and validated in many studies.

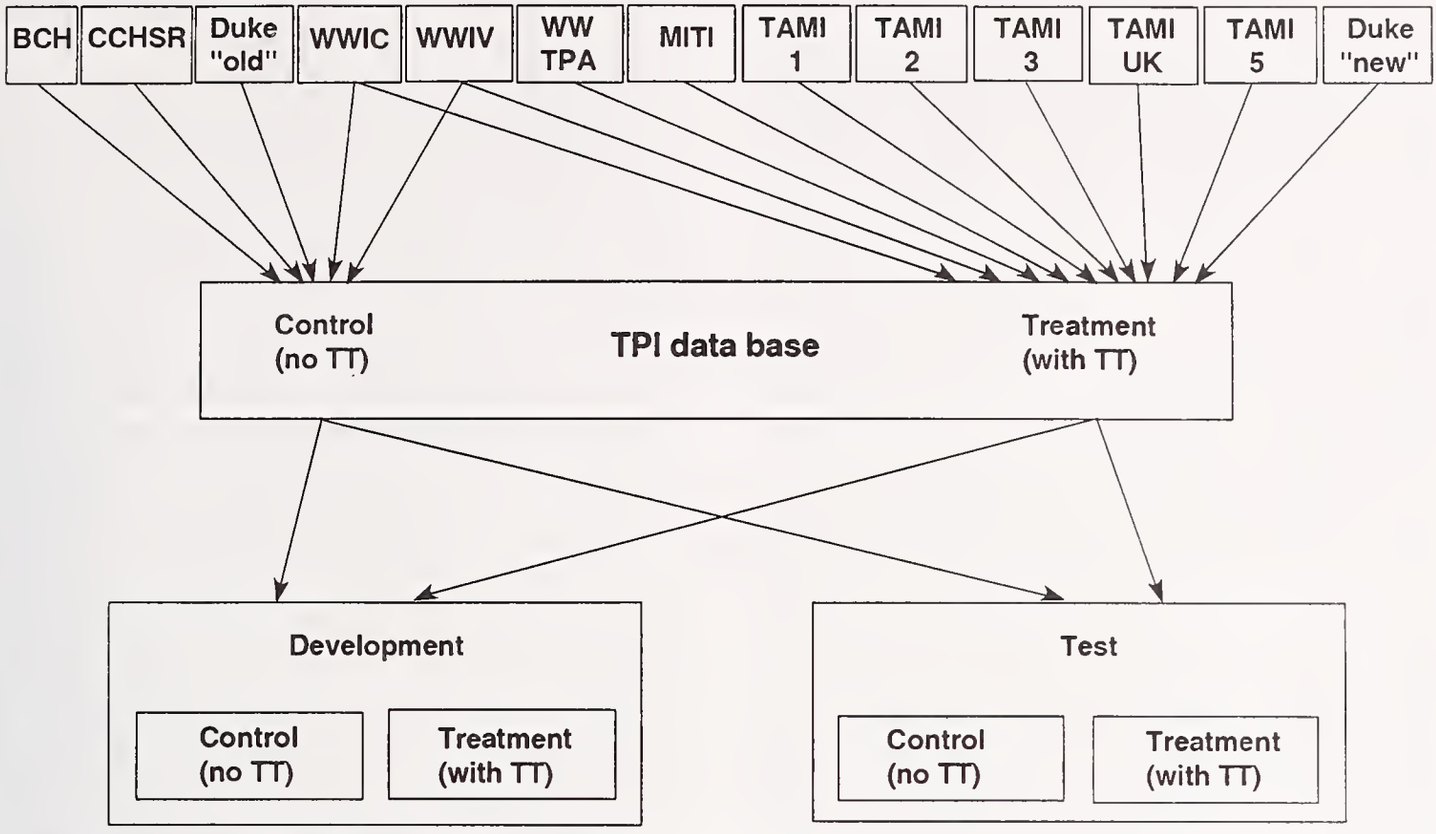
As in other methods of analyzing combined data (such as meta-analysis), the inclusion of only "successful" TT trials does bring a potential for bias. Thus, the inclusion of the MITI registry, the Duke CCU data bank, and the Center for Cardiovascular Health Services Research (CCHSR) data base is critical for their more general data.

As previously indicated and as illustrated in Figure 4, this project includes researchers and data bases from the Western Washington Intracoronary SK Trial, the Western Washington IV SK Trial, the MITI Project, the TAMI Trials, and the Acute Ischemic Heart Disease (AIHD) Predictive Instrument Trials from the CCHSR and Duke CCU data bank. Coordinated by the CCHSR at New England Medical Center and Tufts University School of Medicine, the project also includes researchers associated with the participating studies at the University of Washington, Duke University, and UCLA-Harbor General Medical Center. Below is a brief description of the data from each of these sites.

CCHSR AIHD Predictive Instrument Trials data base. In the Multicenter AIHD Predictive Instrument Trial (Pozen, D'Agostino, Selker, and others, 1984), the CCHSR prospectively collected presenting and acute followup data on 5,773 patients at six New England hospitals during 1979-1981. This data base has been used for the construction of several predictive instruments, including—and most germane to the TPI project—one for AMI acute mortality. Included in this data set were all consenting male patients age 30 or older and female patients age 40 or older, presenting with chief complaints of chest pain, shortness of breath, upper abdominal pain, or dizziness. Over 200 variables were collected, including the patient's clinical history, family history, ED-presenting symptoms, initial and followup ECGs, information on hospitalization, and mortality. Also used in the TPI project are data collected at the Boston City Hospital (BCH) (which was the pilot study for the Multicenter Predictive Instrument Trial), which used the same inclusion/exclusion criteria; these data include 1,781 patients seen in 1976 through 1978 (Pozen, D'Agostino, Mitchell, and others, 1980). Since these data for both predictive instrument trials were collected before the widespread use of TT and other aggressive early interventions, they provide a baseline for assessing the predictive value of different variables for AMI mortality and severity.

Western Washington Thrombolytic trials data. Researchers at the University of Washington in Seattle collected data on the effects of TT in the Western Washington (Intracoronary and IV SK) Thrombolytic Trials. For the first Western Washington TT trial, data were collected in 1981 and 1982 from 14 hospitals on 250 pa-

Figure 4. Component data sets in the TPI data base



tients. Patients that met the enrollment criteria were randomly separated into a treatment group that received intracoronary SK and a control group that received conventional interventions. In the second Western Washington TT trial, 368 patients meeting the enrollment criteria from 27 hospitals for 1983 to 1986 were randomly assigned to either a group that was given IV SK or a control group given conventional treatment. An additional trial included in the TPI data base from this same effort is the Western Washington tPA Trial, which included 160 patients.

Inclusion criteria were age 75 or less, symptoms of AMI with a clear time of onset and typical chest pain for at least 20 minutes, and ST segment elevation in at least two ECG leads (0.15 mV in leads V1, V2, and V3, and 0.1 mV in other leads). Exclusion criteria were: presence of severe disease in other organ systems likely to reduce longevity or the ability to cooperate or give informed consent; prior SK therapy; any contraindication to anticoagulation, such as active ulcer disease, diabetic retinopathy, hypertension with diastolic pressure greater than 120mm Hg; recent surgical procedure (within 2 weeks) or neurosurgical procedure or stroke (within 4 weeks); trauma likely to result in bleeding; presence of central venous lines placed in incompressible entry sites; and prior coronary bypass surgery.

In addition to extensive clinical data to make baseline comparisons between experimental and control groups, the data also contained several valuable outcome comparisons, including mortality, creatine kinase levels, thallium-201 image infarct size determination, clinical functional level, and LVEF as determined by angiography and/or radionuclide scanning. Also collected were AMI location, types of treatment delivered, diagnosis, complications, outcomes, and the time after onset of AMI of initial intervention. This data base also has 1-year and longer followup data.

Myocardial Infarction Triage and Intervention Project registry data. Researchers at the University of Washington are currently gathering data in the MITI Trial (Weaver, Eisenberg, Martin, and others, 1990). In a controlled prospective clinical trial, patients are being given IV tPA either at the ED or in the field by Seattle's well-developed EMS system or in the ED as is now conventional. Preliminary to the intervention phase of MITI, a general registry of all patients receiving TT in their EMS system was developed. Data are being collected on the patients' history and presenting symptoms, type of treatment delivered, LVEF, medications, and outcomes. The data from this TT registry are included in the TPI data base. Because the MITI registry contains information on patients who received TT outside of a

clinical trial, thus reflecting the current usage and general effectiveness of TT, it is an important part of the TPI data base.

Thrombolysis and Angioplasty in Myocardial Infarction Trials data. Also included in the TPI data base are data bases from the TAMI Trials (Bates, Califf, Stack, and others, 1989; Topol, Bates, Walton, and others, 1987; Topol, Califf, Kereiakes, and George, 1987; Topol, Califf, George, and others, 1988; Topol, Fung, Kline, and others, 1986). All patients in these trials received some form of TT. Data collected include presenting symptoms, presenting and followup ECGs, diagnosis, treatment, location of MI, time since symptom onset to TT intervention, complications, and data on the performance of other revascularization procedures.

An additional feature of the TAMI data is that special interventions—most importantly PTCA—were done in different branches of the TAMI studies for comparison to just IV TT (Topol, Califf, George, and others, 1987). This will allow us to examine the impact of emergent additional invasive procedures in the use of TT and thus provide the possibility of having several treatment options within the recommendation to treat with TT. As combinations of TT and other procedures are made more commonly, this feature of the TPI would be desirable. The decision on whether this feature will be included will depend on modeling considerations (whether reliable PTCA primary and interaction terms can be created) and the opinions of the senior clinicians in this study.

Duke CCU data bank. The Duke CCU data bank is a general clinical registry of all patients admitted to the Duke Coronary Care Unit since 1969. Data gathered include patients' clinical history, medications, interventions during hospitalization, and long-term followup. The TPI project uses two subsets of this data base: one, those patients admitted between 1976 and 1984—in the prethrombolytic era—who did not receive TT and two, those patients admitted in 1986 or later who received TT. Patients admitted in the transition year of 1985 were not used in this project because of uncertainty of the TT use practices of different physicians and the potential treatment selection biases. As with the CCHSR and MITI registry data bases, because these data were not collected as part of a controlled clinical trial, they enhance our ability to study the outcomes of usual use of TT, that is, the effectiveness of TT.

Step 2. Development of a collaborative investigative team. Critical to the combining of data from many clinical trials is the very active involvement at all levels from each of the collaborating sites. The need for compatible data in a project like this, and a clear understanding of the situations under which the data were collected, requires continual input from each site. Moreover, beyond the special data sets that the participants bring, their expertise and experience in their areas are critical to the optimal conduct of the project.

To develop and maintain a collaborative team in the TPI project and to tap complementary areas of expertise, we have used quarterly TPI investigator meetings (rotating among sites), weekly fax updates, and weekly conference calls among principal investigators at all sites. All these modes of communication and cooperative endeavors develop and maintain the trust and "investedness" that are critical to the success of this kind of work.

Step 3. Preliminary determination of the availability and characteristics of data in each data base. As a foundation for consideration of specific variables and analyses, data from component data bases had to be carefully described as to the availability and characteristics of their candidate variables. Before the first TPI project investigators meeting, each site sent to the coordinating center a copy of all data codebooks and data specifications (such as coding, availability, and how variables were obtained), specific study protocols, and inclusion and exclusion criteria used at each site. A detailed summary of the variables and data available in each data base was compiled in a large workbook (Figure 5) and iteratively circulated back to the individual sites. The careful itemizing of each variable, inclusion/exclusion criteria, and protocol details was a critical first step in developing an understanding of the data at each site and the potential confounding factors present in the different data bases. Also included in this step was the determination of the practicality of collecting additional data at each site to supplement the data bases, such as by returning to the medical records and use of the National Death Index (NDI).

Step 4. Specification of the dependent variables (predicted outcomes). One of the first tasks of the TPI project investigators was the specification of the outcome variables. Since a predictive model can be no better than the data and variables on which it is based, the optimal specification of the outcome variables is critical. The final forms of the TPI outcome variables were selected based on considerations of the clinical importance of the variable, its measurement characteristics, and the prevalence and distribution of the variable.

During early planning of the TPI project, it was envisioned that the TPI would consist of three component models predicting hospital mortality, change in LVEF, and complications from the use of TT. However, based on preliminary reviews of data from each site and meetings between all collaborators, several important changes were made in the initial list of outcomes and their forms. These considerations, which are typical of

Figure 5. An example of variable availability in each component of the TPI data base

SYSTOLIC BLOOD PRESSURE

	MCHSRU	WWIC	WWIV	WWTPA	MITI	TAMI1	TAMI2	TAMI3	TAMI4	TAMI5	TPI
VARIABLE	SYSBP	ERSBP	ERSBP	ERSBP		SYSBP	SYSBP	SYSBP	SYSBP	SYSBP	SYSBP
Definition	INITIAL SBP IN ER	INITIAL SBP AT ENTRY TO HOSP	INITIAL SBP AT ENTRY TO HOSP	HIGHEST SBP PRIOR TO TX		BASELINE SBP	BASELINE SBP	BASELINE SBP	BASELINE SBP	BASELINE SBP	
Special Characteristics											
Format / Values	CONT 0-270	CONT 0-240	CONT 60-230	CONT 80-220		CONT 60-220	CONT 80-200	CONT 70-206	CONT 60-193	CONT 50-220	
Comments		SOME OF THE MISSING VALUES ARE BEING OBTAINED BY CHART REVIEW (MISSING = 58)	THERE IS ALSO AN "INITSBP" DEFINED AS "FIRST" BLOOD PRESSURE THIS MAY HAVE BEEN TAKEN IN AMBULANCE			BASELINE SBP IS DEFINED AS FIRST RECORDED SBP SINCE ENROLLMENT	BASELINE SBP IS DEFINED AS FIRST RECORDED SBP SINCE ENROLLMENT	BASELINE SBP IS DEFINED AS FIRST RECORDED SBP SINCE ENROLLMENT	BASELINE SBP IS DEFINED AS FIRST RECORDED SBP SINCE ENROLLMENT	BASELINE SBP IS DEFINED AS FIRST RECORDED SBP SINCE ENROLLMENT	

our approach to variable selection, are outlined below for illustrative purposes.

Our original plan for predicting acute mortality was to predict whether a patient survived the presenting AMI through hospitalization. In fact, originally there was little other choice because longer term followup data were not uniformly available for all subjects in all of the participating studies. Because of variations in the typical lengths of hospitalization at different locations over the period of time covered by the TPI data base, and because of the growing interest by professional groups and Federal agencies in using standard end-point definitions for acute hospitalization mortality, we felt that 30-day mortality as our "acute" mortality end point would be a better outcome to predict.

Although preliminary analyses suggested that hospital and 30-day mortality data were extremely close (and we will model both and investigate any areas of discrepancy for potential biases), it was decided that the advantages of 30-day mortality and the desirability of getting long-term mortality (see below) justified the effort of going to the NDI to supplement data sets not already having complete long-term mortality data (for example, the AIHD Predictive Instrument Trials data base). By doing this, 30-day and 1-year mortality data were made available for 85 percent of the TPI data base, and the remaining missing data are being obtained.

Originally, in the absence of uniform long-term survival data, we planned to use LVEF as a proxy for long-term mortality. Although cardiac global LVEF is an important clinical outcome in its own right, it also is an important short-term proxy outcome for long-term survival. Of the basic cardiac clinical indicators, it is generally the single best predictor of long-term survival for most AMI patients. Thus, when the TPI project was being planned, LVEF was selected as one of the outcomes. Beyond its long history of use as this kind of proxy, at that time it was widely believed that the benefit of TT was entirely mediated through salvage of myocardium, as reflected by preservation of systolic left ventricular function. Numerous studies have now evaluated the impact of thrombolytic therapy on resting left ventricular systolic function as measured by LVEF (Califf, Woodlief, and Topol, 1990). These studies point to four themes that reduce the importance of LVEF as an adequate surrogate measure for the impact of TT on patient welfare. First, although the impact of TT on global LVEF is measurable, it is relatively modest. Second, systolic global LVEF is not the only, nor probably the best, measure of the impact of TT on myocardial salvage

and patient benefit. Third, LVEF's measurement is fraught with difficulty and potential biases. Frequently, 10-20 percent of patients have technically inadequate studies, and these failures usually occur in patients with the largest infarctions and who thus are capable of showing the greatest change. Further problems result from the dropout of patients due to death or medical complications. Fourth, and most importantly, there is now mounting evidence of a substantial dissociation between the reduction in mortality and LVEF. The large mortality trials have demonstrated a greater mortality reduction than would be predicted by measured changes in LVEF (Van de Werf and Arnold, 1981). Also, in trials that measured both LVEF and survival, differences have usually been observed in opposite directions (Ohman and Califf, 1990). Further, improvement in nonfatal outcomes has been measured in trials where no difference in global LVEF has been apparent (Califf, Topol, and Stack, 1991). These paradoxical findings have led to evidence that much of the benefit of TT is mediated through a reduction in sudden death, improved healing and remodeling of the ventricle, and improved collateral flow mediated through arterial patency, rather than by improved systolic function (Braunwald, 1989; Ohman and Califf, 1990).

These recent findings suggest that LVEF alone is not an optimal proxy for the long-term impact of TT and thus not an optimal outcome variable for the TPI. When, as described above, we found that we would be able to obtain true long-term mortality data on TPI project subjects by use of the NDI, we came to two conclusions about the use of LVEF for the TPI: one, LVEF as a predicted outcome for the TPI should be supplanted by 1-year mortality, and two, in place of LVEF, the usefulness of the TPI would be enhanced if an expected size of infarction based on ECG, QRS, and ST scoring could be provided as non-TT-contingent information for the emergency clinician.

All of the major trials of TT have now demonstrated continued maintenance of the benefit of TT on mortality out to 1 year of followup (GISSI, 1987; Ohman and Califf, 1990), and 1-year mortality is clearly a very meaningful outcome for patients and one by which treatment effectiveness might be judged. Thus, we were very pleased to be able to add this outcome to the TPI. Although LVEF will no longer be included as a directly predicted TPI outcome, we continue to feel that the presentation of information that reflects the patient's acute pathophysiologic state would be very useful for the clinician treating TT candidates. Thus, the TPI will now have an ECG-based estimate of infarct size based on the Selvester QRS score. Since Selvester and Laks (members of the TPI project investigator team) have already incorporated the QRS score into a computerized electrocardiograph, this component of the TPI is quite feasible.

Discussions of the TPI outcome variables also led to the issue of cardiac arrest, a particular outcome related to death that has particular importance for AMI. Indeed, because of its clinical importance and the potential impact of TT, sudden cardiac arrest was added as an outcome to the TPI. Because this variable was not explicitly defined in most of the component data bases, however, the decision to add this new outcome included the decision to obtain primary medical records for all patients who died or had cardioversion or defibrillation. (This is being done as part of a case-control study described in Step 8.) Because of the importance of this outcome, the TPI investigators felt that the additional effort—consisting of a major study including acquiring supplementary data via primary medical record review of its own with over 60 primary variables and 800 observations from over 100 hospitals—was justified.

Initial discussions about our plans to predict "serious complications of TT" led to the conclusion that the most serious and feared complication, hemorrhagic stroke, should be predicted directly. Thus, it was decided to take a similar approach to that used for cardiac arrest to obtain data for the stroke component of the TPI. To do this, the medical records of all patients who may have had a hemorrhagic stroke also were reviewed and additional data collected (again, as part of the additional data collection efforts described in Step 8). The collected data include copies of computerized tomography (CT) scans, magnetic resonance imaging, and autopsy reports as well as extensive clinical information. For each patient, these data will be reviewed by a committee of experts to accurately determine the nature of the stroke. The use of records from all of the component TPI data bases will result in detailed information on over 50 TT-treated patients who had a stroke, thus making it the largest such data base of TT-related strokes currently available.

The above-described changes in our original outcome variable plans for the TPI illustrate the importance of this step. Specifically, based on new research done since the initial planning of the TPI project and on the added availability of long-term mortality data primarily via the NDI (as well as a willingness to return to medical records to obtain selective important data), the TPI's predicted outcomes have been greatly enhanced. Thus, now the TPI will have component models comparing the impact of the use of TT on 30-day mortality, 1-year mortality, cardiac arrest, strokes, and serious bleeds and an additional model providing an ECG-based estimate of infarct size.

Step 5. Determination and specification of the independent variables. During this step, a candidate pool of independent variables was created and the exact form of each variable determined. At the onset of the TPI project, a list was sent to each site asking the investigators to list all variables that might be of importance in predicting or affecting the selected outcomes. Each site generated two lists: one, a list of all variables that could potentially be important for the TPI (regardless of the practical availability of the data), and second, a list of variables that were considered necessary for creating successful models of TT outcomes. These lists were combined and circulated for additional comments. Through this iterative process among all TPI investigators, a collection of potential variables for the TPI data base was created. This list was further refined at the TPI investigators' meetings, where the discussions included the clinical importance of each of the variables as a predictor of one or more of the outcomes, as well as their implications for measuring the impact of TT (based on new and ongoing research).

At these meetings, the status within each data base component of every potential independent variable was discussed. For each variable, a code sheet was created that listed specific information from each component study about that variable (see example in Figure 5). This included how the data were obtained, additional related data, how the variable was coded, and what effect, if any, the relevant study inclusion/exclusion criteria might have on the variable. The detailed information exchanged at these meetings by investigators, statisticians, programmers, and others concerning the variables was critical for an understanding of each of the data bases. The insights into each of the component data bases went far beyond what would have been otherwise possible. The information required for combining variables from separate data sets to create the specific variable form that would be used in the overall TPI data base (for example, how exactly data for a given variable were acquired, when and where the data were obtained, and the exact formulation of the variable) could not have been obtained simply from the study codebooks and references. Finally, based on issues of scientific merit and practical considerations and in a highly iterative process, a final list of candidate independent variables for the TPI data base was created and the coding format specified.

When the key variables were considered, it was clear that some had insufficient detail or were absent from some or all data sets. The former problem occurred with ECG data, for which it quickly became evident that we needed more detail and uniformity in ECG variables than were available through the different data bases. Therefore, as outlined in Step 8, we went back to the medical records of all subjects and obtained a copy of the presenting and followup ECGs, which we then had uniformly interpreted and coded in great detail for each of all 12 leads. An example of the latter problem arose in the cases of the predictive models for stroke and cardiac arrest; as with the dependent variables, sufficiently uniform and detailed independent variables were not available in the participating data sets. Again, as outlined in Step 8, this led to a return to the primary medical records.

The additional data collection efforts were not taken lightly; they added substantially to the project's required effort and expense. Our structured process of considering each variable made it clear when this was necessary, however, and these efforts greatly enhanced the utility of the TPI data base. The importance of this and the prior steps in the process were reinforced in the decisionmaking to undertake additional data coding and acquisition. This was possible primarily because of the consensus of all participants that was built in the specification of key variables.

Step 6. Specification of inclusion/exclusion criteria for treatment and control groups. When using data originally collected in separate clinical trials for which inclusion/exclusion criteria differ between trials, specification of uniform criteria for adding data to the combined data base is critical. Of particular concern in the TPI project were systematic differences between the patient populations represented in the TT and non-TT control groups. Differences in these respective populations' risks for predicted outcomes would confound predictions of the benefit of TT. Thus, for data from the non-TT trials and registries (the CCHSR Multicenter, BCH, and Duke CCU data bank), criteria for inclusion in the TPI data base were developed to select those patients who would have been eligible to receive TT in order to create comparable treatment and control subjects.

An example of the approach we used in this step is our handling of the AIHD Predictive Instrument Trials. As outlined earlier, these trials included all patients (males over age 30, females over age 40) who presented with ischemic symptoms (shortness of breath, chest pain, dizziness). To select comparable patients to those in the TPI data base receiving TT, patients in this data set who were over age 75, did not have chest pain as their primary complaint, did not have ST elevation on their initial ECG (as read by either the ED physician or an expert reader), or were discharged home from the ED were excluded from the TPI data base. The remaining patients were included if they had either an ED diagnosis of AMI and were admitted to the CCU or a final diagnosis of AMI and an ED diagnosis of AMI, rule/out AMI. To-

gether these criteria created a group of control patients who would have qualified for TT treatment and inclusion into the participating TT trials. Based on these criteria, 396 patients from the CCHSR data set were included in the TPI data base. Although the hospital death rate of patients in the various subsets of the data was not used in establishing the above criteria, the final acute hospital death rate of the patients from the CCHSR selected into the TPI data base exactly matched the rate of the control patients in the Western Washington IV Trial (8.6 percent), supporting the validity of the use of these criteria. Analogous, but generally lesser, adjustments were made in the combining of other data sets.

Step 7. Addition of new data sets to the TPI project. For the same reasons that led us to use multiple data bases to create the TPI data base, we also continuously review completed and ongoing TT and related trials for possible additional inclusion of these studies. Within the constraints of logistic and resource limitations, we have considered for inclusion new data sets that: one, significantly strengthen a critical component of the project, or two, provide a useful test set for the constructed models.

The first example of adding new data sets actually occurred at the very start of the project. Since, as formulated in our original plan, 80 percent of the data in the TPI data base were originally collected on patients who received TT, limitations to the power of detecting significant variables and on the construction of the TPI models were primarily due to the size of the group of patients who did not receive TT. Thus, it was decided to use the patients enrolled in the Duke CCU data bank before 1985 who did not receive TT (patients admitted after 1985 who received TT were admitted to the treatment group). Additionally, the BCH AIHD Predictive Instrument Trial data base was also added at this point. The inclusion of these additional patients greatly increased the ability of our group to model accurately the factors relating to the predicted outcomes on patients who do not receive TT and thus would increase the power and accuracy of the final TPI models. These data sets also increase the racial and ethnic diversity within the TPI data base. Several additional data sets have been acquired for the independent testing of the TPI in Phase Three. These include other major TT trials and data collected on patients before arrival at a hospital by EMS.

Step 8. Acquisition of new primary data. As indicated in the steps related to specifying the dependent and independent variables (Steps 4 and 5), in order to create a data set that would optimally support the development of the TPI models, we decided to acquire additional primary data from medical records. The reasons for needing these additional primary data included:

1. Missing data on some patients for variables critical to the models;

2. Inadequate detail and/or uniformity of important data; and

3. New data (variables) required for modeling.

As indicated earlier and described below, the TPI project required additional data for each of these reasons. Because of the size of the TPI data base and the number of hospitals involved (127), the acquisition of additional primary data in the TPI project significantly increased the time and effort required for creation of the combined data base. To minimize the number of trips to each hospital and the amount of work required, a customized data collection form (Figure 6) for each patient was constructed to obtain all (and only) the required information on that patient. Nonetheless, although we believe this approach is extremely valuable for the construction of combined data sets—as illustrated in our examples—clearly this kind of effort must be justified and planned prior to use in any project such as the TPI.

Despite the large size of the TPI data base, the issue of loss of subjects because of missing variables was a problem: 15 percent were found to be missing primary variables. The loss of this number of subjects from the modeling process, as well as the potential for bias introduced by the loss of these subjects, led to the decision to return to records to get missing key information. Thus, for information such as whether the patient had a coronary artery bypass graft, time since symptom onset to hospitalization or treatment, initial systolic blood pressure, ECG data, and outcome, medical records were reviewed. Because of the number of hospitals involved in the TPI data base, this data collection was done only for subjects from hospitals (79) at which there were cardiac arrest or stroke subjects as well. Overall we expect to have over 90-percent complete data on the primary variables after the entry of these additional data.

The need for additional primary data collection to address the problem of inadequate detail or uniformity is best illustrated in the case of the ECG data in the TPI data base. Measurements from the ECG are the primary predictors of the impact of TT and thus will be critical components of the TPI models. It was clear from the initial review of the different coding systems for ECG data in each of the component data sets, that the ECG data would have to be re-recorded into a uniform system. Additionally, it became clear that a more detailed measurement and coding scheme—including detailed measurements on key components, such as the ST segments in 0.5-mm increments, for every lead in each ECG—would greatly improve the usefulness of this

Figure 6. Electrocardiogram data coding form

Patient name_____ ECG category _____ Admission
Study name_____ _____ Day
Study no:_____
ECG date:_____ Time _____
(24 hour clock)

QRS duration:_____msec
Rhythm: Sinus_____ Atrial flutter - Fib._____ Other_____

Confounding factors: 0 = None; 1 = LVH Sokolow-Lyon, Cornell Voltage, Romhilt-Estes (circle two); 2 = RVH Butler-Leggett; 3 = LBBB; 4 = RBBB; 5 = LAHB; 6 = LPHB; 7 =PACED; 8 =WPW

Prior ECG MI: 0 = No; A = Anterior; I = Inferior; P = Posterior; L = Lateral; AP = Apical

Prior % QRS MI _____(QRS score x3 in leads without acute epicardial injury)
(on admission ECG only)

Acute MI location: 0 = No; A = Anterior; I = Inferior; P = Posterior; L = Lateral; AP = Apical

Acute ST changes: ST deviation at J point (+ or - to nearest 0.5 mm)				
I_____mm		V1_____mm	V4_____mm
II_____mm	aVL_____mm	V2_____mm	V5_____mm
III_____mm	aVF_____mm	V3_____mm	V6_____mm

ST - predicted % MI_____ (Aldrich)
Anterior 3[1.5 (number leads with ST↑) = 0.4]
Inferior 3[0.6 (ΣST↑ in II, III, aVF) + 2]

Maximal T amplitude (+ or - to nearest 0.5 mm from PR baseline) (check if diphasic)				
I_____mm		V1_____mm	V4_____mm
II_____mm	aVL_____mm	V2_____mm	V5_____mm
III_____mm	aVF_____mm	V3_____mm	V6_____mm

Phase of acute MI: 0 = Don't know; 1 = Hyperacute; 2 = Early acute; 3 = Late acute; 4 = Early subacute; 5 = Late subacute; 6 = Chronic

Abnormal Q waves (msec)					
_____ I ≥ 30	_____aVL ≥ 30	_____V1 any	_____V3 any	_____V5 ≥ 30
_____ II ≥ 30	_____aVF ≥ 30	_____V2 any	_____V4 ≥ 20	_____V6 ≥ 30
_____III ≥40				

Total percent QRS MI _____ (Applicable only after acute epicardial injury has resolved)

critical group of variables. Although this entire process required great effort (including the finding of the patient's initial and followup ECG, reading and entering 40 measurement variables per ECG, and the inevitable required tedious data cleaning), it gives the TPI data base an unparalleled level of ECG detail for many patients and creates a critical foundation for creating the TPI predictive models.

The decision to make TPI models predicting TT-related hemorrhagic stroke and cardiac arrest included the decision to collect entirely new data to create these models. For the stroke model, variables such as neurological signs or symptoms, TT dose, results of blood tests, results of CT scans, information on blood transfusions, and the patient's functional status at discharge needed to be added to the TPI data base. For the cardiac arrest model, variables such as special ECG data, serum electrolytes, exact time of death, the onset and progression of the cardiac arrest, and its relation to receiving TT were also needed. Specifically, 214 variables were added for the stroke analyses and 75 variables for the cardiac arrest. To allow collection of these data within realistic time and resource constraints, we chose to collect data on all study subjects having had a stroke (51), for a fourfold number of stroke control subjects (204), all TPI data base patients with a cardiac arrest (400), and an equal number of cardiac arrest control subjects (400).[6] Thus, for each of these 1,055 subjects—along with the patients for whom missing data were to be collected, for a total of 1,720 subjects—customized data collection forms and supporting documents and manuals were created; a return to 79 hospitals was required in order to complete the TPI data base. As a general strategy in combining clinical data bases, we feel that although such an effort should not be undertaken lightly, in cases where key analyses depend on such variables, this approach is well-justified.

Step 9. Checking for consistency of effects across the TPI data base. Despite efforts preceding this step aimed at making the data from the different contributing data bases consistent, some method is needed to confirm that this has been successful and to detect any previously undetected problems. We believe that this is best done by checking for consistency of important effects among the contributing data bases across the entire TPI data base.

Although it is expected (and desired) that the contributing data bases will represent different populations and procedures to be legitimate parts of the same data set, we should expect consistency in well-recognized effects, such as the increased risk for mortality related to advanced age, low blood pressure, and size of infarction. We fully expect some variation in the impact of treatment in different data bases, and it is explicitly not the exploration of the TT treatment effect that this step involves. Rather, this step's task is to detect unexpected differences in fundamental effects—that is, differences in fundamental risk factors for outcomes or their magnitudes—that might reveal an underlying lack of comparability in variables or the presence of unidentified confounding effects.

Although good ancillary information is essential to any useful data base, the activities of this step relied on a particularly high level of data base documentation. Thus, with every iteration of data reformatting and recoding of variables from different data bases to match the TPI data base, all changes were documented and communicated to the sites, along with summary statistics for verification. In addition to a TPI data base codebook, an "annotated" codebook was created, listing along with the final coding, related information for each variable from the original data bases (for example, original variable name, coding, data restrictions, and related variables). This kind of information and the involvement of personnel at all levels from each site were critical to checking the consistency of effects.

This checking process also depended on having very carefully cleaned data, requiring repeated detailed data checks and close collaboration of the coordinating and contributing sites. How the lack of clean data could confound a primary effect is illustrated by the variable, time since symptom onset to the performance of the initial (treatment) ECG.[7] To do this, the recorded dates of the re-read ECGs were checked against the enrollment data in the main data base to insure that the first ECG was in fact the one being used. Where discrepancies were found, correct ECGs were obtained, coded, and added to the data base.

After cleaning and documentation were done, one further issue was addressed before the checking of effects was performed. Given our plan to keep part of the TPI data base as an independent test set for checking the predictive models (see Step 10), we considered the issue that some of this step's activities could be seen as pre-

[6]The decision to obtain additional primary data from a review of medical records and charts also changed the anticipated analysis and model construction for these outcomes. Because of the large number of observations present in the TPI data base, a review of all 4,000 charts was impossible, and thus a case-control design was developed to analyze each outcome. Blocking on hospital and trial, a one-on-one sample of patients who did not have a cardiac arrest and a four-to-one sample of patients who received TT and did not have a stroke and patients who received TT and had a stroke were selected.

[7]As outlined early in this paper, and shown in Figure 2, time from symptom onset and treatment is a dominant influence on the impact of TT.

liminary study analyses on the planned test data, as opposed to mere preparation of the data base. Aside from the critical role this activity plays in the construction of the merged data base, several features convinced us of the legitimacy of this approach: the analyses for this step were only descriptive and univariate—using basic forms of selected variables—for generally accepted effects and did not include tests of statistical significance or comparisons between TT-treated and control subjects of the impact of TT. Beyond these features that differentiate this activity from our plans for creating the TPI models, we also kept these analyses to the very minimum necessary to establish basic data set compatibility.

Specifically, our approach included four comparisons of the data between:

1. The treatment groups of the two (IV and intracoronary) Western Washington TT trials to detect any confounding influence due to differences in their modes of TT administration;

2. Patients in the treatment group of the Western Washington SK Trial (who received IV SK) and patients in the TAMI trials who received IV TT (tPA) without additional invasive procedures;

3. The Western Washington SK Trial treatment group and patients in the MITI registry who received TT (tPA) in the ED; and

4. Patients in the control groups of both Western Washington trials (who received no TT) and AMI patients from the CCHSR and Duke CCU data sets.[8]

For the comparison between data sets involving patients who did not receive TT, we compared the effects of presenting clinical features known to predict mortality (such as hypotension and high Killip class) and ECG variables known to be predictive of AMI mortality (such as anterior Q waves and extensive ST segment elevation) on the outcome mortality.

Step 10. Division of the data base into development and test data sets. The central reason for combining the component data bases was to provide sufficient numbers of subjects to generate valid detailed models. This requirement includes having sufficient numbers to allow validation of predictive models on a separate "test" portion of the data set (Selker, Griffith, and D'Agostino, 1991a, 1991b). Thus, the TPI data base was split into a development data set, on which the component TPI models will be constructed, and a test data set that will be used for model validation.

The approach used in splitting the TPI data base was to ensure sufficient numbers for developing a detailed model as well as for testing the accuracy of predictions. Although equal-sized data sets would seem reasonable, the data requirements for the former are greater than for the latter, which led us to put two-thirds of the data into the development set and the remaining one-third into the test set. Patients were assigned randomly to either set, blocked only on study of origin, to insure consistency between the sets of the proportions of patients from any one component data base.

After this process was completed, Phase Two modeling activities could begin. These activities, along with the subsequent phases of the TPI project, are briefly described below.

Phase Two: Construction of the TPI

For each TPI outcome, we will construct, analyze, and test a separate predictive model. For inclusion in each initial model, from the initial pool of clinically significant predictive variables, we will select a promising subcollection. Based on each variable's clinical importance, its overlap and interactions with other variables, and its contribution to model performance, variables will be progressively winnowed down, reformatted as necessary, and finally included in the initial predictive models. These models will then be analyzed by conventional criteria (such as goodness-of-fit tests, receiver-operating characteristic [ROC] curve areas, and R-square) and also by measures directly related to their likely clinical performance, including ability to discriminate between clinically meaningful groups, "calibration" of predicted likelihoods of outcomes throughout their range of predictions, and the types of misclassification errors made. Potential new clinically important variables will be added to those retained, and models will be reformulated, as needed, to optimize performance for projected clinical use.

Related to the issues raised in the discussion of Step 10 is the need during modeling to look for potential confounding influences of the component data sets. Thus, throughout development, models will be checked to determine if the observed effects are "parallel" across the appropriate groups from each component data base. This will be done in four ways, by:

1. Looking at interaction terms in the logistic regressions between selected variables and variables identifying individual data bases (this will be done separately for treatment and control patients);

[8]The inclusion of the TAMI patients receiving TT plus other treatments—for example PTCA—will be based on the acceptability of the inclusion of the TAMI patients who received only TT. If the latter group is analogous to those receiving TT in the Washington and MITI trials, then presumably differences seen in the extra intervention group will be legitimately compared with the combined standard TT treatment group for modeling purposes.

2. Carefully analyzing model residuals across sites to determine subsets of the data that are not accurately reflected by the selected predictive variables;

3. Using standard influence diagnostics to see if individual sites have subsets that are making the model unstable; and

4. Doing repeated modeling on all but one of the component data sets and then testing model performance on the remaining data set.

Phase Three: Testing of the TPI on Independent Data Sets

In Phase Three, we will test the models developed in Phase Two on independent data sets. These tests are intended to: one, validate the construction and performance of the TPI and two, validate the generalizability of the models. Although these tests are important for any predictive model, because of the complexity of the TPI models, the implications of their use, and the inclusion of many sources of data in the TPI data base, this type of testing is particularly critical in this case.

Each of the predictive models constructed on the development portion of the TPI data base will be assessed on the test portion set aside at the end of Phase One. The models' performance on these new data will be checked using the same conventional and clinical performance-related measures used in Phase Two. In addition, the performance will be checked on individual sites. New variable weights will be determined on the test set, and these coefficients will be carefully compared with the original model to check individual variables' contributions. These steps should provide an important assessment of the validity of the TPI models.

A limitation of the use of the TPI data base test set is that, having been selected from the same data sets as the data used to construct the models, this validation does not directly address the issue of the models' generalizability to other sites. Thus, we will also test our models on data from clinical trials that were not included in the TPI data base. This will allow us to validate the TPI models' performance on data entirely unrelated to their construction and to test their potential applicability to a broad variety of potential ED and EMS settings. This checking of the generalizability of the TPI will be a critical step towards its desired eventual widespread use.

Phase Four: Incorporation of the TPI into a Computerized Electrocardiograph for Real-Time Clinical Use

Once the TPI has been developed and tested sufficiently to demonstrate its potential use and safety as a clinical tool, we will implement it in the form that should be most clinically useful—that is, incorporation into a computerized self-interpreting electrocardiograph. In the ED and EMS settings, where the ECG is central to the initial evaluation and decisionmaking for cardiac patients, as done with our AIHD predictive instrument (see Figure 1), the provision of the TPI's predictions as part of the printout at the top of the ECG tracing would seem to be the most desirable form (see Figure 3). The fact that electrocardiographs can now transmit full 12-lead ECGs (and thus the TPI's predictions) via cellular phone links allows for a range of uses of the TPI's output both in the ED and EMS vehicles and at remote sites. Its predictions could range from acting as an instant "second opinion," to being used as part of formal treatment criteria. This flexibility will be critical for the TPI's use in pilot testing and in clinical trials and, after it becomes commonly used, in clinical practice. Nonetheless, however accurate its predictions, it is our intent that—as with our earlier predictive instruments—its output will still be intended to supplement, not replace, clinicians' judgments.

Although at this point in the TPI project we will be focusing primarily on the creation of the clinical tool version of the TPI, we will be looking ahead to the TPI's potential use for other applications. Thus, as part of the programming done during this phase, we will also create microcomputer and hand-held calculator programs of the TPI for potential use in nonclinical situations.

Phase Five: Clinical Trial of the TPI

Assuming that the TPI performs as desired in the previous phases, Phase Five will be a prospective clinical trial of the impact of the use of the TPI on the administration of TT and the impact on patients' survival and complication rates. Although the methods described in this paper for combining clinical data sets and for the construction of predictive models are powerful and useful, before it is put into general clinical use, the actual impact of the final TPI will need to be tested in a clinical prospective randomized clinical trial.

Conclusion

The creation of a "super data base" from data previously collected for major clinical trials and registries is a powerful method for constructing models and testing hypotheses. If done with care at each step, this technique allows for:

1. A greater number of subjects to be studied than would be possible in a single study;

2. The efficient use of a major research resource, namely carefully acquired and validated clinical data; and

3. The construction of models that may otherwise be impossible for ethical or practical reasons.

The 10-step approach to combining clinical data bases outlined in this paper (see the table), which constitutes Phase One of the TPI project, has proved practical and deserves further exploration. We believe that it should have general utility for medical effectiveness and clinical research on issues for which single extant data bases are insufficient and/or the collection of new data is not feasible.

As detailed at the outset of this paper, in clinical use the TPI will include a set of detailed outcome predictive models that will provide predictions of TT's likely effectiveness for a given patient for use by the clinician in the real-time emergency setting. The TPI's models will presumably also have application to more typical, retrospective, analyses of the effectiveness of this medical treatment. We believe, however, that bringing this kind of outcomes-based analysis to the real-time treatment of AMI might be of considerable benefit beyond that to be expected from more conventional effectiveness research approaches. At least for a given patient in the midst of an AMI, the TPI's provision of information about expected treatment effectiveness for prospective use should have more impact than a retrospective effectiveness analysis. In bringing this approach to hospital EDs and to the homes and streets of communities via EMS, we believe this approach has the potential to "improve the effectiveness of effectiveness research."

References

American Heart Association. (1976). *Coronary risk handbook: Estimating the risk of coronary heart disease in daily practice.* New York: Author.

Applebaum, D., Weiss, A.T., Koren, G., and others. (1986). Feasibility of pre-hospital fibrinolytic therapy in acute myocardial infarction. *American Journal of Emergency Medicine*, 4, 201-204.

Bates, E.R., Califf, R.M., Stack, R.S., and others. (1989). Thrombolysis and angioplasty in myocardial infarction (TAMI-1) trial: Influence of infarct location on arterial patency, left ventricular function, and mortality. *Journal of the American College of Cardiology*, 13, 12-18.

Bergman, S.R., Lerch, R.A., Fox, K.A.A., and others. (1982). Temporal dependence of beneficial effects of coronary thrombolysis characterized by positron tomography. *American Journal of Medicine*, 573-581.

Braunwald, E. (1989). Myocardial reperfusion limitation of infarct size, reduction of left ventricular dysfunction, and improved survival — Should the paradigm be expanded? *Circulation*, 79, 441-444.

Califf, R.M., Topol, E.J., Stack, R.S., and others. (1991). An evaluation of combination thrombolytic therapy and timing of cardiac catheterization in acute myocardial infarction: The TAMI 5 randomization trial. *Circulation*, 83, 1543-1556.

Califf, R.M., Woodlief, L.H., and Topol, E.J. (1990). Left ventricular ejection fraction may not be useful as an end point of thrombolytic therapy comparative trials. *Circulation*, 85, 5.

Chesebro, J.H., Knatterud, G., Roberts, R., and others (1987). Thrombolysis in myocardial infraction (TIMI) trial, phase I: A comparison between intravenous tissue plasminogen activator and intravenous streptokinase. *Circulation*, 76, 142-154.

Collen, D., Topol, E.J., Tifenbrunn, A.J., and others. (1984). Coronary thrombolysis with recombinant human tissue-type plasminogen activator: A prospective, randomized, placebo-controlled trial. *Circulation*, 77, 1012-1017.

Franzosi, M.G., Mauri, F., Pampallona, S., and others. (1987). The GISSI study: Further analysis. *Circulation*, 76, 1152-1156.

Grim, P., Feldman, T., Martin, M., and others. (1987). Cellular telephone transmission of 12-lead electrocardiograms from ambulance to hospital. *American Journal of Cardiology*, 60, 715-720.

Grines, C.L., Nissen, S.E., Booth, D.C., and others. (1990). A prospective, randomized trial comparing combination half-dose t-PA with streptokinase to full dose t-PA in acute myocardial infarction: Preliminary report. *Journal of the American College of Cardiology*, 15, 4A.

Gruppo Italiano per lo Studio Della Streptochinasi nell'Infarcto Miocardico (GISSI). (1986). Effectiveness of intravenous thrombolytic therapy in acute myocardial infarction. *Lancet*, 1, 397-401.

Gruppo Italiano per lo Studio Della Streptochinasi nell'Infarcto Miocardico (GISSI). (1987). Long-term effects of intravenous thrombolysis in acute myocardial infarction: Final report of the GISSI study. *Lancet*, 2, 871-874.

Gruppo Italiano per lo Studio Della Streptochinasi nell'Infarcto Miocardico (GISSI). (1990). GISSI-2: A factorial randomised trial of alteplase versus streptokinase and heparin versus no heparin among 12,490 patients with acute myocardial infarction. *Lancet*, 336, 65-71.

Hartmann, J.R., McKeever, L.M., Bufalino, V.B., and others. (1986). Intravenous streptokinase in acute myocardial infarction: Experience of community hospitals served by paramedics. *American Heart Journal*, 111, 1030-1034.

Institute of Medicine. (1990). *Report of a study: Acute myocardial infarction: Setting priorities for effectiveness research* (pp. 1-48). Washington: National Academy of Sciences.

ISAM Study Group. (1986). A prospective trial of intravenous streptokinase in acute myocardial infarction (ISAM): Mortality, morbidity and infarct size at 21 days. *New England Journal of Medicine*, 314, 1465-1471.

ISIS Steering Committee. (1987). Intravenous streptokinase given within 0-4 hrs of onset of acute myocardial infarction reduced mortality in ISIS-2. *Lancet*, 1, 502.

Kennedy, J.W., Martin, G.V., Davis, K.B., and others. (1988). The Western Washington intravenous streptokinase in acute myocardial infarction randomized trial. *Circulation*, 77, 345-352.

Kennedy, J.W., Ritchie, J.L., Davis, K.B., and Fritz, J.K. (1983). Western Washington randomized trial of intracoronary streptokinase in acute myocardial infarction. *New England Journal of Medicine*, 309, 1477-1482.

Kircher, B.J., Topol, E.J., O'Neill, W.W., and Pitt, B. (1987). Prediction of infarct coronary artery recanalization after intravenous thrombolytic therapy. *American Journal of Cardiology*, 59, 513-515.

Marder, V.J. and Sherry, S. (1988). Thrombolytic therapy: Current status (Pt II). *New England Journal of Medicine*, 318, 1585-1595.

Norris, R.M., Barnaby, P.F., Brandt, P.W.T., and others. (1984). Prognosis after recovery from first acute myocardial infarction: Determinants of reinfarction and sudden death. *American Journal of Cardiology*, 53, 408-413.

Ohman, E.M. and Califf, R.M. (1990). Thrombolytic therapy: Overview of clinical trials. *Coronary Artery Disease*, 1, 23-33.

Pozen, M.W., D'Agostino, R.B., Mitchell, J.B., and others. (1980). The usefulness of a predictive instrument to reduce inappropriate admissions to the coronary care unit. *Annals of Internal Medicine*, 92, 238-242.

Pozen, M.W., D'Agostino, R.B., Selker, H.P., and others. (1984). A predictive instrument to improve coronary-care-unit admission practices in acute ischemic heart disease. *New England Journal of Medicine*, 310, 1273-1278.

Selker, H.P., D'Agostino, R.B., and Laks, M.M. (1988). A predictive instrument for acute ischemic heart disease to improve coronary care unit admission practices: A potential on-line tool in a computerized electrocardiograph. *Journal of Electrocardiology*, S11-S17.

Selker, H.P., Griffith, J.L., and D'Agostino, R.B. (1991a). A tool for judging coronary care unit admission appropriateness, valid for both real-time and retrospective use. *Medical Care*, 29, 610-627.

Selker, H.P., Griffith, J.L., and D'Agostino, R.B. (1991b). A time-insensitive predictive instrument for acute myocardial infarction mortality: A multicenter study. *Medical Care*, 29, 1196-1211.

Simoons, M.L., Serruys, P.W., Van Den Brand, M., and others. (1986). Early thrombolysis in acute myocardial infarction: Limitation of infarct size and improved survival. *Journal of the American College of Cardiology*, 7, 717-728.

Topol, E.J., Bates, E.R., Walton, J.A., and others. (1987). Community hospital administration of intravenous tissue plasminogen activator in acute myocardial infarction: Improved timing, thrombolytic efficacy and ventricular function. *Journal of the American College of Cardiology*, 10, 1173-1177.

Topol, E.J., Califf, R.M., Kereiakes, D.J., and George, B.S. (1987). Thrombolysis and angioplasty in myocardial infarction (TAMI) trial. *Journal of the American College of Cardiology*, 10(5), 65B-74B.

Topol, E.J., Califf, R.M., George, B.S., and others. (1987). A multicenter randomized trial of intravenous recombinant tissue plasminogen activator and emergency coronary angioplasty in acute myocardial infarction. *New England Journal of Medicine*, 317, 581-588.

Topol, E.J., Califf, R.M., George, B.S., and others. (1988). Coronary arterial thrombolysis with combined infusion of recombinant tissue-type plasminogen activator and urokinase in patients with acute myocardial infarction. *Circulation*, 77, 1100-1107.

Topol, E.J., Fung, A.Y., Kline, E., and others. (1986). Safety of helicopter transport and out-of-hospital intravenous fibrinolytic therapy in patients with evolving myocardial infarction. *Catheterization and Cardiovascular Diagnosis*, 12, 151-155.

Van de Werf, F. and Arnold, A.E.R. (1981). Discrepancies between the effects of coronary reperfusion survival and left ventricular function. *Lancet*, 1, 1367-1368.

Van de Werf, F., Ludbrook, P.A., Bergmann, S.R., and others. (1984). Coronary thrombolysis with tissue-type plasminogen activator in patients with evolving acute myocardial infarction. *New England Journal of Medicine*, 310, 609-612.

Vermeer, F., Simoons, M.L., Bar, F.W., and others. (1986). Which patients benefit most from early thrombolytic therapy with intracoronary streptokinase? *Circulation*, 74, 1379-1389.

Verstraete, M., Bernard, R., Bory, M., and others. (1985). Randomized trial of intravenous recombinant tissue-type plasminogen activator versus intravenous streptokinase in acute myocardial infarction: Report from the European Cooperative Study Group for Recombinant Tissue-type Plasminogen Activator. *Lancet*, 842-847.

Villemant, D., Barriot, P., Riou, B., and others. (1987). Achievement of thrombolysis at home in cases of acute myocardial infarction. *Lancet*, 1, 228-229.

Weaver, W.D., Eisenberg, M.S., Martin, J.S., and others. (1990). Myocardial infarction, triage and intervention project: Phase 1 patient characteristics and feasibility of prehospital initiation of thrombolytic therapy. *Journal of the American College of Cardiology*, 15, 925-931.

White, H.D., Norris, R.M., Brown, M.A., and others. (1987). Effect of intravenous streptokinase on left ventricular function and early survival after acute myocardial infarction. *New England Journal of Medicine*, 317, 850-855.

Combining Data from Different Studies

Mark A. Hlatky, M.D.

In epidemiologic and clinical studies, there are substantial incentives to combine patient data from several centers. One major reason is that a single center may not have enough patients with a given illness to have adequate statistical power to test hypotheses of interest. Combining data from several centers may provide sufficient numbers of patients. A second reason to combine data is to obtain a broader range of subjects with a disease—for example, by including more patients of a certain age or gender or with mild or severe forms of disease or simply including different geographic locations and practice styles.

Multicenter studies—particularly multicenter controlled randomized clinical trials—are a natural outgrowth of these efforts. The investigators in a multicenter study will spend considerable time and energy establishing a common method of collecting data and developing standard definitions and data forms. Central laboratories are often established to assure uniform analysis of key measures in the study. These elaborate maneuvers are designed to permit data collected at the different sites to be combined and analyzed. Prospective multicenter trials are difficult, time-consuming, and expensive. As a result, these studies are not always feasible. Consequently, investigators are often interested in combining data that were not collected according to a common protocol.

Meta-Analysis

Meta-analysis has been embraced recently as a technique for combining findings from different studies to obtain an overview of a clinical problem (L'Abbe, Detsky, and O'Rourke, 1987; Sacks, Berrier, Reitman, and others, 1987). The use of meta-analysis has been fueled by the recognition that many clinical trials are too small to provide definitive evidence of the value of a particular therapy (Freiman, Chalmers, Smith, and Kuebler, 1978). In meta-analysis, the researcher gathers all the clinical trials on a given topic and abstracts information about outcome in each study. Although a variety of statistical methods have been used, a central feature of these techniques is that the results of the different studies are pooled, not the primary data.

Meta-analysis of observational studies has not been used as often as meta-analysis of randomized trials (Steinberg, Thaker, Smith, and others, 1991). Meta-analyses of observational data, such as epidemiologic investigations, present analytic and practical challenges beyond those encountered in analysis of randomized trials. Single prognostic factors can be assessed most easily using meta-analysis. Multivariable analyses cannot be readily combined using the standard statistical techniques of meta-analysis. Apart from any technical difficulties in the analysis of epidemiologic data is the problem of bias. Randomized trials can assure that treated and control groups are established without bias, so that combining unbiased studies should lead to an unbiased result. If the initial data are observational in nature and subject to bias, combining data does not remove the bias. Meta-analysis of epidemiologic studies is, therefore, inherently less reliable than meta-analysis of randomized trials.

Pooling Primary Data

One can go beyond pooling results to pooling data. This approach must face the problems that arise from differences in the methods used to collect data in different studies. Pooling of primary data from different epidemiologic studies has not been attempted very often. One notable previous effort was the Pooling Project (The Pooling Project Research Group, 1978), a consortium of eight different cohort studies of cardiovascular disease: Albany Cardiovascular Health Study, Chicago Peoples Gas Company Study, Chicago Western Electric Study, Framingham Heart Disease Epidemiology Study, Los Angeles Heart Study, Minnesota Business and Professional Men Study, Minnesota Railroad Workers Study, and Tecumseh Heart Study. After initial dis-

Dr. Hlatky is Chief of the Division of Health Services Research and Associate Professor at the Stanford University School of Medicine.

cussions, three of these population groups were not included: the Los Angeles cohort lacked information about smoking and had different measurement methods for blood pressure and cholesterol; the Minnesota Business and Professional Men were volunteers, not population samples; and the Minnesota Railroad Workers had a relatively brief followup period. It is notable that measurement methods for cardiovascular risk factors are relatively simple and were well-standardized at the time of the Pooling Project. Data concerning electrocardiograms (ECGs) were the most difficult to combine, but coding of the ECGs in all studies was done using a standard method (the Minnesota Code); the technicians in each study had all been trained in Minnesota, assuring that uniform methods were used in each of the studies. Thus, the Pooling Project was successful but was built on a foundation of prior methodological collaboration among cardiovascular epidemiologists.

More recently, the effect of blood pressure on cardiovascular mortality was examined by pooling primary data from nine prospective cohort studies (MacMahon, Peto, Cutler, and others, 1990). This investigation included three of the five groups in the original Pooling Project. The study examined only a single independent variable (blood pressure) that is measured in a relatively standard fashion. Data from each of the individual studies were displayed separately and then combined to provide more precise estimates of the quantitative relationship between diastolic blood pressure and outcome. Neither the Pooling Project nor the more recent collaboration on blood pressure explicitly addressed heterogeneity among data sources, differences in measurement technique, or secular trends in outcome.

Measurement Issues for Pooled Data

These two instances of primary data pooling are remarkable successes and highlight the tremendous difficulties in pooling primary data. The greatest challenge is reconciling data collected by different investigators in different ways. Variables of interest may have been collected at some but not all sites, and therefore it may be impossible to assess differences between centers, much less adjust for them. For example, in the prethrombolytic era, the length of symptoms of acute myocardial infarction prior to presentation to the hospital was rarely recorded, and even if it was, little attention was paid to its accuracy. It is impossible to reconstruct this information retrospectively, so the effects of time to presentation or patient physical findings, electrocardiographic signs, and laboratory findings are impossible to assess in older data sets.

A difficult problem in comparing data from different centers is that the same pathophysiologic concept (for example, myocardial ischemia) may be measured using different methods in different centers. In the case of acute myocardial infarction, the ECG is of central interest. Different centers often have different standards for diagnosis of "myocardial infarction" or "myocardial injury," so that a data base of ECG interpretations from different centers may be incommensurate. This problem may be addressed somewhat by comparing the written diagnostic criteria at different centers and attempting to map the codes onto a common scale. The ideal solution is to reinterpret the original tracings in a standardized way.

Even if the written standards of data collection and interpretation are identical, the actual performance may differ among centers. Biochemical tests at different centers (for example, cholesterol, creatine phosphokinase) may be performed using the same nominal methods but yield different numerical results. It is possible in principle to align the results from different centers on a common scale by sending aliquots from the same specimen to each center for analysis, but in retrospective studies, one can never be certain that laboratory practices have been stable over time. In multicenter prospective trials, the investigators may decide to use a central laboratory to analyze specimens from all clinical sites, but this option is not available to retrospective studies pooling data between studies.

The problems due to different measurement techniques pose a greater threat to validity than do differences in the patient populations. In general, a diversity of patients is appropriate and desirable to assure a wide spectrum of disease and to blunt any bias due to referral patterns at single institutions. Characteristics of the patients relevant to outcome can be measured and used in statistical models of outcome. The key assumption is that the important determinants of outcome have been measured and that unmeasured patient characteristics have little bearing on outcome. Cardiovascular disease has been closely studied for years, and the delineation of the patient characteristics predictive of outcome has advanced as far as any area within medicine. Nevertheless, not all the important factors affecting prognosis are either understood or measured. Time-related trends are of particular concern in heart disease incidence, mortality, and treatment. General standards of care continue to evolve, so that the outcome of similar patients presenting to the same institution and receiving the "same therapy" may be quite different at different times. Coronary care in 1990 is probably better (and certainly different) from coronary care in 1980, which in turn was different from coronary care in 1970. A carefully performed study at Duke University showed that the results of coronary bypass surgery have improved steadily over a

15-year period (Pryor, Harrell, Rankin, and others, 1987). These changes were attributed to known, but unmeasured, incremental improvements in surgical, anesthetic, and postoperative management over the period of study. A factor related to calendar time can be included in the statistical model in an attempt to account for these changes, but this approach is clearly less satisfactory than direct measurement of the changes in care leading to differences in outcome. Thus, comparisons of patient cohorts assembled at different times may be complicated by secular trends in outcome.

Analytic Issues for Pooled Data

Combinations of primary data from different sources can distort the relationship between variables and outcome. The greatest problems arise when the populations or rates of outcome are substantially different between sites. Table 1 summarizes hypothetical data to show how a spurious relationship may arise between a variable and an outcome measure. Hospital A and Hospital B are assumed to have equal results in old and young patients, but the two hospitals are assumed to differ in their overall mortality and in the proportions of old and young patients. Since the hospital with the higher mortality also has more old patients, a spurious association between age and outcome can be created when the data are pooled.

Conversely, pooling may mask a true relationship between a prognostic factor and outcome. In Table 2, the older patients have a worse prognosis than younger patients in both Hospital C and Hospital D. The two hospitals have different overall mortalities and different proportions of old and young patients, and in this instance, a true relationship between age and outcome is hidden.

Table 1. Hypothetical combination of data creating a spurious correlation of age with mortality

	Old		Young	
	Number	Percent	Number	Percent
Hospital A	40/400	10	10/100	10
Hospital B	5/100	5	20/400	5
Total	45/500	9	30/500	6

Table 2. Hypothetical combination of data masking a true correlation of age with mortality

	Old		Young	
	Number	Percent	Number	Percent
Hospital C	40/400	10	5/100	5
Hospital D	25/100	25	60/400	15
Total	65/500	13	65/500	13

These examples also suggest that one can guard against misleading results by examining the relationship between risk factors and outcome in each individual center. The possibility of serious confounding of outcome by a center can arise if there are wide disparities in the rate of outcome between different sites, differences in the prevalence of risk factors between different sites, or both.

It is helpful to use data plots and graphs to visualize the relationships among variables at different centers. Tests for heterogeneity among centers should be used with the recognition that the statistical power of such tests is low. Stratified analyses can help avoid problems arising from differences among centers.

Advantages of Pooling Data

Despite the considerable methodological difficulties involved in pooling primary data, there can be considerable advantages over using only information available from the literature. Investigators from each of the studies contributing data should be actively involved in a pooling project. The simple act of collaboration on a common project promotes the sharing of insights and experience, as well as data; furthermore, other spinoff projects and collaborations may result. The discussion of measurement issues, which must be resolved in any pooling project, can add to improved—or at least standardized—techniques. The value of increased statistical power and increased breadth of the clinical spectrum of illness have already been cited as strong advantages of data pooling. A less obvious advantage is the chance to perform parallel analyses in independent data sets, since apparent differences among studies can result simply from differences in choice of analytic methods, not from differences in the relationship among variables.

Pooling projects are difficult enterprises, requiring considerable investment of time and effort by the collaborating investigators. There also are considerable methodological limitations to performing pooling studies. Despite these limitations, well-performed pooling projects may advance knowledge in a field in a variety of ways.

References

Freiman, J.A., Chalmers, T.C., Smith, H., Jr., and Kuebler, R.R. (1978). The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: Survey of 71 "negative" trials. *New England Journal of Medicine*, 299, 690-694.

L'Abbe, K.A., Detsky, A.S., and O'Rourke, K. (1987). Meta-analysis in clinical research. *Annals of Internal Medicine*, 107, 224-233.

MacMahon, S., Peto, R., Cutler, J., and others. (1990). Blood pressure, stroke, and coronary heart disease: Part 1, prolonged differences in blood pressure: Prospective observational studies corrected for the regression dilution bias. *Lancet,* 335, 765-774.

Pryor, D.B., Harrell, F.E., Jr., Rankin, J.S., and others. (1987). The changing survival benefits of coronary revascularization over time. *Circulation*, 76 (Suppl. V), 13-21.

Sacks, H.S., Berrier, J., Reitman, D., and others. (1987). Meta-analyses of randomized controlled trials. *New England Journal of Medicine*, 316, 450-455.

Steinberg, K.K., Thaker, S.B., Smith, S.J., and others. (1991). A meta-analysis of the effect of estrogen replacement therapy on the risk of breast cancer. *Journal of the American Medical Association*, 265, 1985-1990.

The Pooling Project Research Group. (1978). Relationship of blood pressure, serum cholesterol, smoking habit, relative weight and ECG abnormalities to incidence of major coronary events: Final report of the Pooling Project. *Journal of Chronic Diseases*, 31, 201-306.

Research on Outcomes of Care Using a Medicaid AIDS Data Base: Overcoming the Curses of Imperfect Data

Barbara J. Turner, M.D., M.S.Ed.
Leona E. Markson, Sc.D.
Robert Houchens, Ph.D.
Thomas Fanning, Ph.D.

Introduction

The depth and precision of primary data continue to serve as the cornerstone of health services research. Increasingly, however, researchers are taking advantage of the breadth and efficiency of research on secondary data. Most of the Agency for Health Care Policy and Research's (AHCPR's) innovative Patient Outcomes Research Teams (PORTs) are capitalizing on these complementary characteristics by conducting analyses on both types of data (Raskin and Maklan, 1991). Creative use of secondary data bases can be an important component of medical effectiveness evaluations, but the task of dealing with secondary data base limitations can be formidable. It involves the recognition of sources of error and deficiencies and the generation of valid, yet practical, solutions for these problems. Since the future of research on medical effectiveness and outcomes of care is likely to retain the model of the PORTs through the use of both types of data, it is essential that the health services research community share approaches for building on the unique strengths of secondary data bases.

In order to explore outcomes of ambulatory care for adults and children with the acquired immunodeficiency syndrome (AIDS) in New York State, our research team uses a secondary data base derived from Medicaid claims and eligibility files. The project does not involve primary data collection; however, we are conducting a small data validation project and augmenting the Medicaid data base with information from a health care provider questionnaire survey and from Government health statistics. We believe secondary data bases can be used to explore many important issues related to medical effectiveness.

This paper provides background information on the outcomes of ambulatory care for AIDS patients, with a particular focus on the study of New York State's Medicaid AIDS population; discusses the components of our research data base as well as some of the weaknesses inherent to most claims-based research; reviews the use of Medicaid data for outcomes research and specific problems; presents the elements of a new severity measure developed for use with longitudinal clinical histories of AIDS patients; and summarizes our approach to defining AIDS risk groups. In addition, the results of some comparative survival and patterns of care analyses conducted on our data will be described, and some conclusions regarding our collaborative effort and suggestions for other teams to follow in this area of health services research will be presented.

Background

To date, the focus of most health services research on AIDS has been on the quantification of its economic impact (Hellinger, 1990; Scitovsky and Rice, 1988). Another important but relatively unexplored avenue of investigation is analysis of patterns of provider care and associated outcomes for AIDS patients. An increasing proportion of the AIDS population is composed of indigent adults and children who are likely to be Medicaid enrollees. These patients may have limited access to services and inadequate continuity of care.

Our research team is conducting an AHCPR-funded project to examine the relationships between patterns of

Dr. Turner is Clinical Associate Professor of Medicine and Dr. Markson is Research Assistant Professor of Medicine, Thomas Jefferson University, Division of General Internal Medicine and Center for Research in Medical Education and Health Care. Dr. Houchens is Senior Statistician, SysteMetrics/McGraw-Hill. Dr. Fanning is Director of Research SURS Systems Management Reports, New York State Department of Social Services.

outpatient provider care, inpatient/emergency room use, and patient survival for AIDS patients of all ages who are enrolled in the New York State Medicaid program. Five groups of Medicaid enrollees with AIDS are identified based on risk group, gender, and age: homosexual/bisexual men, drug-using men, nondrug-using (heterosexual contact) women, drug-using women, and children. We are comparing the patterns of care received by the groups of adults with AIDS.[1] In addition, we are examining provision of care to the rapidly growing pediatric population afflicted with AIDS which is just now becoming a serious health care delivery challenge to cities across the country. The specific objectives of this project are to:

1. Define continuity and types of outpatient provider care for different risk groups of adult AIDS patients and determine specific patterns of care that produce greater than expected health care resource use demands or prematurely poor clinical outcomes; and

2. Evaluate the impact of patterns of provider care and informal caregivers (foster or family/friend) on the clinical course of the disease and on health care resource demands of pediatric AIDS patients.

The subject of outcomes of care for individuals with AIDS may seem somewhat incongruous, since by definition, they have experienced serious complications that generally carry a poor prognosis. Yet the human immunodeficiency virus (HIV) epidemic and disease course of individuals with AIDS are rapidly changing. We now know that infected persons may experience up to a decade or more of relatively minimal symptoms before developing serious complications, and that many late complications can be treated or at least ameliorated (Sande and Volberding, 1988). Thus, HIV infection is increasingly considered a chronic disease that requires close monitoring of antiviral therapies (Volberding, Lagokos, Koch, and others, 1990), early prophylaxis, and treatment of complications. Although it is vital that quality of care for persons with earlier phases of the HIV infection be assessed, it is difficult to identify these individuals in secondary data bases because of "labeling"—reluctance to report HIV infection on claims—and because many individuals do not have the disease diagnosed until more advanced complications occur.

Thus, our research focuses on the population that meets the Centers for Disease Control's (CDC) most recent AIDS surveillance case definition (CDC's, 1987). These patients can be identified relatively easily in our data base, and in recent years, a significant increase in survival rates has been observed for those who develop AIDS-related complications (Lemp, Hirozawa, Araneta, and others, 1991). Differences in clinical outcomes for treatable complications and quality of care have now assumed greater importance even for those diagnosed with AIDS. Indeed, some research has uncovered quality-of-care differences. For example, hospital experience in treating AIDS patients has been found to be related to death rates for *Pneumocystis carinii* pneumonia (PCP) (Ball, Farley, and Turner, 1990; Bennett, Garfinkle, Greenfield, and others, 1989).

Outcomes of care for AIDS is of increasing concern, in part because the demographic profile of the AIDS population is dramatically changing. The more affluent and politically powerful homosexual population was most affected in the early years of the epidemic, but recent studies have found a leveling off in the numbers of seroconverting homosexual/bisexual men (Lindan, Rutherford, Payne, and others, 1989; Winkelstein, Wiley, Padian, and others, 1988). Now AIDS increasingly involves drug users and poor, minority women and children because of a failure to stem the spread of the disease in these groups. In addition, the most recent revision of the AIDS surveillance case definition incorporates more late disease manifestations characteristic of these groups, such as extrapulmonary tuberculosis (Zaccharelli, Pezzotti, Salmaso, and others, 1991).

Infected drug users, women, and children are likely to lack the means to pay for care, resulting in dependence on publicly supported health care programs, such as Medicaid. Unlike homosexual groups, they have not developed powerful advocacy groups to mobilize their communities' resources to help meet their needs. As a result, the AIDS population, like other indigent populations dependent on State and federally funded health care programs, may have restricted access to quality care (Bravemen, Oliva, Grisham-Miller, and others, 1989; Lurie, Ward, Shapiro, and others, 1986; Sullivan, 1989).

The size of this indigent, HIV-infected population is likely to be substantial. Estimates indicate that nearly a half million drug users in the United States are at high risk for both contracting and spreading the disease (Friedland and Klein, 1987). Since 1988, new AIDS cases in drug users in New York City outnumbered those in homosexual/bisexual men (New York State Department of Health, 1990). In certain urban areas, the seroprevalence is reaching substantial proportions; 6 percent of all patients seeking care in an urban hospital in Baltimore were found to be seropositive (Kelen,

[1]For example, homosexual men are likely to "spend down" to Medicaid after they have incurred sufficient out-of-pocket expenses to be eligible for Medicaid enrollment and may have established more consistent sources of outpatient care. Drug users and women are primarily indigent and potentially medically underserved.

DiGiovanna, Bisson, and others, 1989). In addition, the numbers of children developing AIDS are rising at an alarming rate (CDC, 1991).

Provision of adequate care to this substantial and growing population is an issue of high social and ethical concern. Persons infected with HIV may not be able to find a provider for many reasons. Surveys of physicians found many to have serious reservations about treating HIV-infected patients (Smith, Godeau, and Katner, 1989; Thurn, Campbell, and Henry, 1989), and some stated that it is acceptable to refuse treatment to these patients (Gallop, Taerk, Lancee, and others, 1989; Smith, Godeau, and Katner, 1989). Aside from fears about becoming infected, avoidance of AIDS patients may also result from the significant homophobia and aversion to contact with drug users that has been documented in studies of physicians' attitudes (Hall, Knickman, Cross-Durham, and Wittenberg, 1989). Because Medicaid reimbursement rates are generally lower than private insurance rates, physicians and hospitals caring for AIDS patients receive relatively low rates of compensation for their efforts (Andrulis, Weslowski, and Gage, 1989). Thus, there is the added problem of a financial disincentive for physicians, hospitals, and other service providers to care for these patients. In view of these problems, very little is known about current patterns of care and associated outcomes for AIDS patients.

Although some preliminary work has been conducted on the costs and outcomes of care for adults with AIDS, virtually none has been performed on the pediatric population. Only a handful of studies have estimated the basic costs of caring for these children; these studies have been institution-specific and examined small numbers of children (Conviser, Grant, and Coye, 1991; Hegarty, Abrams, Hutchinson, and others, 1988; Kemper and Forsyth, 1988). Preliminary data show inpatient costs for children to be similar to those incurred by adults (Thaul and Ball, 1989), but more detailed research is needed on patterns and outcomes of care for these youngest sufferers of the disease. Children who develop AIDS may be particularly vulnerable to receiving inadequate care due to unstable home environments and the consequences of poverty, drug use, and the effects of HIV infection on their mothers or other family members. Evidence from research on other pediatric conditions suggests that costs of caring for children who have an informal caregiver outside of the hospital are significantly lower than for those children who board in the hospital (Hegarty, Abrams, Hutchinson, and others, 1988; Kemper and Forsyth, 1988). It is unclear, however, whether children remaining in the hospital have better or worse clinical outcomes than those living at home or in foster care. Although pediatric AIDS patients are often treated in tertiary care centers, they can also be cared for by providers with less experience in AIDS care. Consequently, the impact of these patterns of care on patient outcomes is of particular interest to health planners.

As AIDS becomes a less rapidly fatal disease, the major site of care is shifting away from the hospital to the outpatient setting. This shift means that physicians in outpatient settings will be handling a greater proportion of the care for AIDS patients. It is expected that continuity of outpatient care will be an important factor in determining AIDS prognosis and the level of demand for inpatient or emergency services. For example, subpopulations of AIDS patients lacking a consistent source of care may die earlier than expected. Consequently, high priority might be placed on devising ways to link these patients to consistent and experienced ambulatory care settings. Other analyses on outcomes of patterns of care may reveal areas where health care workforce development and training could be more effectively targeted. These research questions need to be addressed with data bases that capture longitudinal histories on sufficient numbers of AIDS cases known to be at risk of having deficiencies in care. As the next section will document, Medicaid files are a potential source of these data.

Significance and Use of Medicaid Data Bases for AIDS Outcomes Research

Medicaid and outcomes research on AIDS and other diseases. Medicaid programs are bearing an increasing proportion of the medical care costs associated with AIDS (Green and Arno, 1990). The percentage of all AIDS patients in New York who were enrolled in the State Medicaid program at some point after diagnosis was reported to have increased from 50 percent in 1983 to 83 percent in 1988 (Fanning, Cosler, Gallagher, and others, 1991). This burden is due both to the rapid spread of the disease in indigent populations and to the personal financial drain for health care expenses incurred by patients with AIDS.

Indigent women and children, even before the HIV epidemic, generally become eligible for Medicaid through the Aid for Families with Dependent Children (AFDC) program (Roper, 1987). In 1986-1987, the majority of pediatric AIDS discharges from a national sample of hospitals had Medicaid coverage (Conviser, Grant, and Coye, 1991). Of the 164 women currently enrolled in the NIH-sponsored Women and Infant Transmission Study (WITS), 81 percent are on Medicaid (WITS Study Group, 1991). Even those who are financially able to afford private insurance will usually qualify for Medicaid at the onset of serious HIV-related

complications because of huge out-of-pocket outlays for care.

When research on outcomes of care is conducted, it is important to remember that AIDS is still a relatively rare condition, with only 215,000 U.S. cases reported since the beginning of the epidemic (CDC, 1992). Claims data represent one of only a few broad-based and cost-effective data sources that capture a large number of patients linked to detailed information on their diagnostic conditions and provider types over the course of the disease. Therefore, health services researchers have good reason to look to Medicaid data bases to explore many issues related to care for AIDS.

Evaluation of health care delivery for those enrolled in Medicaid is not a new focus in health services research, but in recent years, many more studies have analyzed physician participation in Medicaid and access to mainstream care for all types of Medicaid patients (Hohlen, Manheim, Fleming, and others, 1990; Mitchell, 1983; Mitchell and Schurman, 1984; Perloff, Kletke, and Neckerman, 1986). Recognition of the value of research with such administrative data records grew substantially in the past (Blumberg, 1986). Several groups examined the impact of various Medicaid program initiatives on patient outcomes (Carey, Weis, and Homer, 1990; Piper and Ray, 1990). These studies focused on the types of providers treating Medicaid patients and on numbers of services obtained (for example, preventive health care). In addition, Medicaid data were shown to be feasible for monitoring the safety of marketed drugs (Gerstman, Lundin, Stadel, and Faisch, 1990).

Studies examining outcomes of patterns of care often link other data sources—such as birth certificate data—to Medicaid Management Information System (MMIS) files (Schramm, 1986). This approach augments the data on MMIS files with important demographic information on individual Medicaid enrollees. This methodology depends on access to individual identifiers for Medicaid recipients that are linked to other data sources. Use of patient identifiers represents a difficult problem for researchers studying AIDS because of strict patient confidentiality regulations to protect the identities of those infected with HIV. In general, only certain individuals specifically sanctioned to have access to these data (for example, selected employees of the Medicaid program) can link data sets by patient identifiers. Thus, researchers outside of State governments may use Medicaid data bases to study AIDS care, but encrypted identifiers restrict their supplementing Medicaid files with information from other sources.

Our approach to dealing with these issues has been to form a multi-institutional research team, combining staff of the New York State Department of Social Services with researchers and clinicians from other institutions. This team has allowed us to combine expertise on New York's Medicaid program with complementary clinical knowledge and research skills. The following section discusses data base development for our research project and the approaches that have been used to augment our secondary data base while strictly adhering to the patient confidentiality regulations.

Development of Medicaid longitudinal AIDS file. Dr. Fanning and colleagues from the New York State Department of Social Services created a longitudinal Medicaid claims data base comprising nearly 8 million claims for approximately 16,000 AIDS patients treated in 1983-1988. These files are being expanded through 1990. Construction of this data base included:

1. Multiple processing steps to identify AIDS patients;
2. Selection of all claims for these recipients;
3. Linkage of claims under a unique identifier, correcting for multiple identification numbers that occasionally are assigned to Medicaid enrollees; and
4. Inclusion of demographic data from eligibility files.

The New York Medicaid data base is of particular interest for AIDS researchers because the highest proportion of reported AIDS patients in the Nation has come from this State (CDC, 1991). In addition, the demographic mix of AIDS cases in New York—with large numbers of women and drug users—is likely to be more representative of the population that will be afflicted with the disease in the future. In contrast, States such as California have proportionately far fewer persons with AIDS from these groups. One of the key challenges is accurately finding AIDS cases from a claims data base. Case finding and other issues related to the construction of a research data base from Medicaid claims are also examined in this section.

AIDS case finding. Developing an AIDS case-finding methodology has been a major hurdle that must be overcome when meaningful health services research on AIDS is conducted with secondary data bases. The case-finding methodology for our project was developed using California Medicaid (Medi-Cal) data and tested against information from the California AIDS surveillance records, a primary data source (Keyes, Andrews, and Mason, 1991). Briefly, medical records specialists assisted the research team in linking all AIDS-defining and a variety of HIV-related conditions to the most appropriate International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) codes. Our ICD-9-CM screen for AIDS incorporated all codes used to specify AIDS including: several immunodeficiency codes (279.x) used before

October 1986 when new ICD-9-CM codes for HIV-related disease were introduced, specific codes for AIDS (042.x) used after 1986, and the code for PCP (136.3). A matrix of various combinations of these codes supplemented with codes for a variety of AIDS-defining conditions (for example, cryptococcal meningitis) was created for various demographic groups. When this screening algorithm was applied to the Medi-Cal data, slightly more than 80 percent of non-drug-using men age 18-50 who were identified in the Medicaid data base as having AIDS were found to have been reported to the AIDS registry. The screen performed less well for other demographic groups; however, it is important to remember that over 90 percent of persons with AIDS in California before 1987 were non-drug-using men so that this was the best group in which to test the performance of the screen. Based on these analyses that used Medi-Cal data, the case-finding protocol appears to function reasonably well.

Without access to AIDS registry data in New York, refinement of the case-finding protocol can only be based on examination of longitudinal histories of ICD-9-CM coded diagnoses on Medicaid files. The investigator can determine if the patient has, over time, any additional clinical clues that he/she has AIDS versus other forms of immunocompromise (for instance, as the result of chemotherapy or autoimmune disease). This process is frustrating because no final confirmation is available, and the investigator relies only on clinical reasonableness of specific diagnoses given all conditions in the patient's history. However, even comparisons with registry-reported cases cannot serve as a gold standard. Registry data on an estimated 15-20 percent of AIDS patients are missing because of lags in case reporting, changes in case definitions, and failures to report cases (Carr, Coppola, and Barry, 1991; Rosenblum, Buehler, Morgan, and others, 1991).

Longitudinal file components. Once AIDS patients are identified, all claims for each of these Medicaid enrollees are extracted from the master Medicaid claims files. This process requires tracking cases through changes in Medicaid enrollment numbers. In our data base, all patient claims were extracted for 3 years before the appearance of an AIDS code (for example, 042.x) or an AIDS-related condition (for example, PCP) and until death or the end of the last fiscal year on the file. When all claims were extracted, eligibility information was obtained for these recipients. Finally, patient demographic and monthly summary files were created.

The following demographic information is included: age, birth date, sex, zip code of home residence, date of Medicaid eligibility, date of first AIDS-defining diagnosis, and date of death from inpatient or nursing home sources. The files contain information on service use (such as hospital stays, outpatient visits, and drug prescriptions) and up to two diagnoses from outpatient service claims. Hospital service data include length of stay, discharge status, and up to five ICD-9-CM codes for conditions specified in the discharge abstract.

In our research project, monthly profiles were created to summarize both inpatient and outpatient utilization patterns for each patient. Monthly profiles existed for all visits to outpatient treatment sites including hospital-based clinics, free-standing clinics, private physicians' offices, emergency rooms, and special treatment settings such as drug treatment programs. These monthly profiles provided a succinct summary of the often large numbers of claims for services to these individuals.

Weaknesses in the Medicaid AIDS research file. A key variable in our outcomes of care research is the patient's survival time after AIDS diagnosis. One of the problems with longitudinal analyses is right-censoring, persons who appear to be alive at the end of records either because they are, in fact, still alive or they are lost to followup. The latter could have occurred because the individual died without a Medicaid record of the death or moved out of the State. Our most reliable death information has come from inpatient or nursing home sources, but increasing numbers of Medicaid patients die in other settings. In addition, patients occasionally leave their residences after being diagnosed with AIDS to return to their families in other regions of the country. Because of these problems, we have found it necessary to seek additional death data from the National Death Index (NDI).

Our data also do not have reliable ethnic information, and we do not have a direct measure of socioeconomic status. The recording of race information on applications for Medicaid enrollment has been found by State audits to have a high rate of error. We have tried to determine a proxy for socioeconomic status. Most Medicaid patients who were eligible prior to their AIDS-defining diagnosis (for example, at least 6 months before AIDS) are likely to come from a similar indigent socioeconomic background and to be roughly at equivalent risk for problems in quality of care. The group likely to be quite different are those who "spend down" to Medicaid after the development of AIDS-related complications. Thus, our analyses do attempt to take into account the timing of Medicaid eligibility in relation to the diagnosis of AIDS as a way to examine socioeconomic status.

Without access to primary data, we do not have valuable laboratory data such as CD4 lymphocyte count, widely recognized to be an important predictor of developing AIDS and correlated with the degree of advance-

ment of the effects of the virus on the immune system (Phillips, Lee, Elford, and others, 1991; Vlahov, Munoz, Solomon, and others, 1991). Lack of this laboratory variable is currently a greater problem for the study of HIV-infected populations before they develop AIDS because these patients generally do not have multiple clinical complications that permit assessment of stage of disease from diagnostic data. In the future, absence of CD4 lymphocyte count data may be a greater problem for the identification and study of AIDS because of proposed changes in the AIDS case definition which include CD4 count criteria. Other laboratory data have been found to be useful in predicting long-term survival of AIDS patients, including albumin and arterial oxygen levels and several hematologic parameters (Justice, Feinstein, and Wells, 1989). Since these predictors are not found in our data, our approach to measuring severity of illness has relied on clinical data determined from analyses of ICD-9-CM codes.

Provider characteristics file. In addition to detailed patient-level data, the Medicaid files can be an important source of information on the providers of health care services for AIDS patients. For our project, a list of all providers treating AIDS patients in New York was developed, including type of provider (private physician or clinic), location (hospital-affiliated or free-standing clinic), and specialty of both clinics and private physicians. The files can also be used to determine each provider's yearly "Medicaid AIDS experience," since this measure may be related to patient outcomes. Since Medicaid files do not have information on patients with other types of insurance, the provider's general AIDS experience cannot be determined. We believe, however, that the Medicaid AIDS experience measure is important because special skills are required of providers to address specific clinical, social, and economic issues common to this population.

To gain a sense of the bias involved in a measure of Medicaid AIDS experience as compared with a measure of total AIDS care experience, we asked Dr. John Colombotos of Columbia University to add to his physician and nurse survey on HIV/AIDS several questions regarding the proportion of all HIV-infected patients seen by respondents in the past year represented by Medicaid HIV-infected patients. These data, along with physician demographic information, will be used to examine the degree to which the Medicaid AIDS experience is associated with the general AIDS experience for specific types of physicians. Admittedly, a measure of provider AIDS care experience is less precise for care delivered in clinics and outpatient departments.

Use of the Medicaid AIDS Longitudinal File for Outcomes Research

A particular strength of the Medicaid data base is in evaluating patterns of care and assessing the impact of care on patient survival and resource use. These research issues might be rephrased in the terminology of the Medical Treatment Effectiveness Program. Do patient survival patterns differ depending on the patient's type of provider? Is the use of health care resources dependent on the type of outpatient care provider? The elements needed to address these questions are largely held within our research file, but certain additional data items are required to increase the comprehensiveness of the data base. As mentioned, the inclusion of NDI data will decrease the right-censoring in our data and increase the accuracy of our estimates of patient survival. Out of the deaths identified on the file through Federal fiscal year 1989, 46.7 percent were from nonhospital reports. This high proportion of cases with death data from less reliable sources necessitates a merger with NDI data.

Administrative data bases were used to obtain estimates of the costs of caring for AIDS patients (Andrulis, Weslowski, and Gage, 1989; Kelly, Ball, and Turner, 1989). Because of extensive service use and charge information on Medicaid claims files, this data source offers broad-based information on the costs of caring for AIDS patients that cuts across all sites of care in one State. Of course, the charges incurred by Medicaid patients in New York cannot be generalized to other payers in the same State nor to Medicaid programs in other States because of the unique fee structures of each State's Medicaid program (Cohen, 1989). Nevertheless, Medicaid data have been used to compare the relative costs of care by gender and risk group (Solomon, Hogan, Bouknight, and Solomon, 1989; Turner, Markson, McKee, and Fanning, 1991).

For analyses of resource utilization, the research files encompass vast amounts of data on service use, but researchers using Medicaid files need to be aware of several possible flaws in the data. First, Medicaid enrollment may be intermittent because of failure to meet eligibility requirements in specific months. Second, individuals who are enrolled in Medicaid after the diagnosis of AIDS will not have their full clinical and service use history on the Medicaid file. These problems can be addressed to some degree.

The problem of intermittent eligibility appears to be relatively important. We examined the proportions of adults with AIDS who were continuously enrolled in

Medicaid for the first 6 months after they were diagnosed with an AIDS-related condition and found that approximately 60 percent had continuous monthly enrollment. Women had somewhat higher proportions of continuous eligibility than men (63 percent versus 59 percent, respectively). These analyses, however, include patients who are right-censored and whose eligibility patterns may be confounded by out-of-hospital death or by leaving the State without notifying Medicaid.

To eliminate right-censoring, we examined the eligibility patterns of patients in the last 6 months of life for those with reliable death dates. In this population, we found higher proportions remaining continuously enrolled for all groups except male non-drug users. In drug users, continuous enrollment was observed for 79 percent of men and 84 percent of women. In non-drug users, continuous enrollment was found for 60 percent of men and 75 percent of women. Thus, the problem of intermittent eligibility for the "spend-down" group still appears to be significant.

It should be noted, however, that patients are usually enrolled in Medicaid when they have a serious disease-related complication requiring frequent evaluations, testing, or inpatient care because their medical expenditures for this care are likely to be sufficiently great to meet enrollment standards. Consequently, episodes of noneligibility probably correspond to episodes of clinical stability, and the clinical data on our files should capture most of these more serious disease-related complications.

Clearly the intricacies of Medicaid eligibility greatly influence the point in the disease course that patients can enroll and the degree to which they can remain eligible. As previously noted, the histories of individuals who "spend down" to Medicaid after the diagnosis of AIDS would be subject to left-censoring; in other words, their history, as reflected in the Medicaid file, would begin sometime after they developed AIDS. Persons with AIDS are one of the "categorical groups" qualifying for Medicaid enrollment, but these individuals also must meet a financial test. New York is one of a group of States with a medically needy program that sets higher income and asset criteria for Medicaid eligibility than is set by the Federal Government for the AFDC or the Supplemental Security Income (SSI) programs (Ellwood, Fanning, and Dodds, 1991). New York has an additional support program for low-income persons who do not meet the Federal criteria for Medicaid assistance. This "Home Relief" program is totally State-funded and uses a less restrictive definition of disability than the SSI program. In Ellwood's study of the patterns of Medicaid eligibility for New York Medicaid AIDS patients, nearly one-third of patients were eligible through this program. Effectively, this means that an HIV-infected individual can develop a serious complication, such as a pneumonia—which would qualify him/her for Home Relief—and then continue on Medicaid enrollment as AIDS-related complications develop. The net effect of these features of Medicaid eligibility in New York is that a higher proportion of HIV-infected persons are enrolled relatively early in the course of the disease and persons with AIDS can "spend down" to the program more rapidly.

Even with more generous Medicaid eligibility criteria, 41 percent of New York Medicaid patients in 1983-1986 were found to have AIDS onset dates prior to their Medicaid eligibility. To address this left-censoring issue, our project will merge data from another source to gain a better understanding of the biases resulting from the absence of this clinical and service use information. Examples of other data that we have identified to address this issue include the New York Statewide Planning and Research Cooperative System Bureau longitudinal hospitalization file, in which all inpatient stays for an individual are linked over time, and the Empire Blue Cross/Blue Shield data base, which is the primary private insurer for inpatient care in New York State. A merger of the Medicaid data and one or the other of these two data bases will serve to address several questions:

- How does the AIDS onset date derived from Medicaid data compare with the date developed from merged data?
- How does the AIDS-defining diagnosis developed from MMIS data alone compare with that constructed from merged data?
- What is the pattern of clinical complications experienced before Medicaid eligibility?
- How do survival curves vary when computed with only Medicaid claims data versus merged data?
- How great are the resources consumed after the onset date but before Medicaid eligibility (for example, the number of hospitalizations for AIDS before Medicaid enrollment)?

We will compare results of each of these analyses by gender, risk group, and eligibility type. Although the answers to these questions will be very informative, the ideal solution to the problem of left-censoring would be to perform all study analyses on an all-payer file, but legal issues obviate passing a merged research file to the entire research team. A research file of multiple linked primary and secondary data bases has been constructed in Maryland, representing an enormous cooperative effort of both public and private agencies (Hidalgo, 1990).

Even with these flaws, Medicaid data hold great promise for researchers interested in AIDS treatment patterns and outcomes. For instance, Medicaid data can be structured to provide a longitudinal profile of patient care that has been delivered by multiple providers (both inpatient and outpatient), and it can be used to track the use of new pharmaceutical agents.

Zidovudine is now recognized to be an important weapon against the HIV virus. In the late 1990s, most infected patients with AIDS-related complications who are not in specific drug trials should be on this drug unless they have a contraindication (for example, severe neutropenia) or they have not responded to it. Since Food and Drug Administration approval, the New York Medicaid program has paid for zidovudine prescriptions, and our records contain relatively comprehensive data on the use of this drug. Analyses conducted by Leon Cosler of the New York State Department of Social Services showed substantial differences in access to this drug. Women and drug users are significantly less likely to receive zidovudine after AIDS diagnosis than are non-drug-using men. Our results support analyses of access to the drug in the Maryland AIDS population (Moore, Hidalgo, Sugland, and Chaisson, 1991). Interestingly, in preliminary analyses made possible by our work on severity measures for AIDS, we found that patients with higher severity AIDS-defining diagnoses were less likely to receive the drug.

Development of Measures for Use in Secondary AIDS Data Bases

The severity index for adults with AIDS. A central component of medical effectiveness research is the ability to account for differences in patient severity of illness. The wide clinical spectrum of complications seen in AIDS and the variable impact of these complications on expected survival make it necessary to develop specific severity of illness measures for AIDS. Because no preexisting severity measure could be used with longitudinal secondary data bases, we created a new survival-based severity index for adults with AIDS (SIAA) based on a construct developed from expert opinion. This construct was empirically tested on a longitudinal data base containing clinical diagnoses and survival information for more than 3,900 adult AIDS patients enrolled between 1983 and 1986 in the New York Medicaid Program and followed through 1988 (Turner, Markson, McKee, and others, 1991). SIAA differs from prior work conducted by these investigators on the severity of a hospitalized AIDS patient (Turner, Kelly, and Ball, 1989) in its applicability to the longitudinal history of an AIDS patient's disease experience.

This index was developed in three phases. First, a six-member expert panel identified clinical factors that influence the survival of AIDS patients, and they estimated survival times (in months) for each of these factors individually and in specified combinations. Second, a model was created to represent the relative severity of each of the conditions ordered by the experts' median survival estimates. Third, the hierarchy of AIDS-related conditions, translated into ICD-9-CM codes, was tested on the Medicaid analysis file. Based on expert consensus, two elements of the patient's disease history—the first AIDS-defining diagnosis and subsequent AIDS-related complications within a fixed time interval after diagnosis—were integrated into the severity index. Testing of these severity hierarchies, described below, resulted in a matrix that groups patients surviving at least 3 months after the diagnosis of AIDS into four categories based on the severity of the first AIDS-defining diagnosis and the presence of the most severe complication occurring within 3 months of diagnosis.

The elements in the severity hierarchy include 19 types of AIDS-defining conditions according to the CDC's latest revision of the AIDS surveillance case definition (CDC, 1987). The experts also identified a list of 80 complications that are likely to influence a patient's prognosis after the diagnosis of AIDS, including all AIDS-defining conditions (for example, cryptococcal meningitis) as well as non-AIDS-defining conditions (for example, gram-negative septicemia). In order to identify the first AIDS-defining illness of each individual, all inpatient and outpatient physician claims were searched for the first appearance of ICD-9-CM code(s) consistent with conditions meeting the CDC's AIDS case definition. Concurrent or subsequent clinical complications, as determined from all claims, were grouped and examined in 3- and 6-month time intervals after the initial AIDS diagnosis. Complications noted during the 6-month interval were consistent with those noted during the 3-month interval, except that the number of patients surviving 6 months was substantially smaller. Therefore, the model examines only complications within a 3-month interval after the occurrence of the first AIDS-defining diagnosis. When two or more AIDS-defining conditions occurred within the same month, patients were grouped by the most severe complication according to the expert panel's severity hierarchy. The severity program, however, records the occurrence of all conditions from the expert panel's list of complications, permitting analyses of frequencies of specific types of complications and combinations of complications.

The task of specifying conditions by ICD-9-CM codes has been one of the most tedious aspects of this project,

since codes sometimes lack sufficient specificity to distinguish separate types of opportunistic infections with differing prognoses (for example, cryptosporidiosis versus isosporiasis) or specific patterns of infections (for example, chronic herpes simplex infection). There are no specific codes for either wasting syndrome or AIDS dementia complex. The clinical reasonableness of probable codes for these and other conditions were evaluated by reviewing a minimum of 10 randomly selected clinical histories for each of the conditions with problematic coding. Medical records specialists were also consulted on coding practices. This review resulted in refinement of codes used to identify complications and development of certain decision rules for classifying diagnoses. For instance, we decided to accept codes for AIDS dementia complex and wasting syndrome only when recorded concurrently with or following the appearance of an immune deficiency code. This approach distinguished AIDS dementia complex from transient mental status changes prior to AIDS and wasting syndrome from AIDS-related complex.

Another well-recognized problem in using a severity measure based on secondary data is that of the accuracy of conditions recorded in these types of data. In terms of the quality of inpatient diagnostic coding, fortunately, AIDS patients tend to be treated in large hospitals that are reported to have lower coding error rates (Demlo, Campbell, and Brown, 1978). However, we have only the first five discharge codes available on our data base, and complex AIDS patients may have some significant conditions that are not recorded among these top five codes.

The codes from outpatient visits have presented an even greater dilemma. Inconsistent coding is a particular problem. For example, patients may have immunodeficiency coded three or four different ways in the space of several months. In addition, some physicians clearly chose one diagnosis that they recorded for a patient, such as AIDS, regardless of what other problems were experienced by the patient. Thus, our ability to detect less severe complications, such as thrush, from outpatient data is somewhat limited.

When physicians bill separately for hospital visits, we have been able to check the validity of the diagnoses provided on these visits with the conditions recorded on the patients' hospital discharge bill. For example, the diagnosis of pneumonia was found occasionally to have been recorded on a physician inpatient visit bill early in the course of a hospitalization, but the hospital discharge codes show that the correct diagnosis after evaluation was PCP, suggesting that the physician's diagnosis was based on preliminary information before the identification of an etiologic agent. We eventually decided to delete clinical information from physicians' inpatient visits from analysis to avoid counting the same condition as two separate ones.

To investigate the quality of the codes analyzed in our study, we are conducting a small data validation project with the cooperation of two groups at Brown University and Mt. Sinai Medical Center. Inpatient and outpatient charts of Medicaid patients who have been treated at several New York State institutions have been abstracted for clinical information that is being correlated with diagnostic information contained on the longitudinal research file. Through this project, we hope to gauge the degree of miscoding in our analysis file. Our expectation is that it will be small in the inpatient setting but more significant in the ambulatory setting.

Despite all of these issues regarding the quality of the clinical information on our analysis file, we have found that the expert panel's severity hierarchies have been largely confirmed by the survival experience of the patients in our Medicaid data base, with survival times decreasing with increasing severity of the AIDS-defining diagnosis and subsequent complications. Table 1 displays the frequency of cases and types of AIDS-defining conditions in the severity hierarchy. Three groups were formed after an iterative process comparing Kaplan-Meier survival curves for these conditions. Severity groups 1 and 3 included persons with multiple conditions who had similar within-group survival curves. Severity group 2 comprised only individuals with PCP because of the large numbers of patients with this condition. The percentages of patients by defining diagnosis severity group are 34, 48, and 17 percent in groups 1, 2, and 3, respectively. Some conditions lead to longer and others to shorter survival times than predicted by the experts, but these variations are few. As a group, the patients in severity group 1 had the longest observed survival times with a median survival of 25 months, as compared with 10- and 7-month median survival times for defining diagnosis groups 2 and 3, respectively.

The median survival times for patients grouped according to the traditional hierarchy of PCP, Kaposi's sarcoma (KS), and all other AIDS-defining conditions were also determined in our population. We found that both the KS group and the "all other" group had the same median survival time, 14 months, in contrast to the 10-month survival for patients with PCP. This classification clearly buries a great deal of diversity of some common AIDS-defining conditions in the "all other"

Table 1. Frequency of AIDS diagnoses by severity group and median survival times –1983-1986 New York State Medicaid program

Severity group	Diagnosis	Median survival	
		Number	Months
1	Kaposi's sarcoma - no organ involvement	245	14
	Chronic mucocutaneous herpes simplex	316	26
	Extra-pulmonary tuberculosis	154	20
	Candida esophagitis	537	18
	Wasting syndrome	220	14
	Salmonella sepsis	31	7
2	*Pneumocystis carinii* pneumonia	2,001	10
3	Disseminated atypical mycobacterial infection	51	8
	Kaposi's sarcoma - organ involvement[1]	9	–
	Cryptococcosis	190	7
	Herpes simplex - organ involvement[1]	16	–
	Disseminated cytomegalovirus infection	85	5
	Cryptosporidiosis/isosporiasis	32	5
	Toxoplasmosis	153	8
	Disseminated histoplasmosis[1]	18	–
	AIDS-associated lymphoma	104	10
	AIDS dementia complex	88	4
	Disseminated coccidioidomycosis[1]	5	–
	Progressive multifocal leukoencephalopathy[1]	9	–

[1]Median survival not estimated due to insufficient sample size.

group. We believe that our three groups provide much more clinical distinction for researchers, if only AIDS-defining diagnosis is available.

We also found that differentiation of survival experiences of AIDS cases could be further improved by examining secondary complications within 3 months of the patient's first AIDS diagnosis. Patients dying within this 3-month interval (21 percent of the population) were excluded. The death rate within 3 months was 9, 27, and 29 percent for cases in defining diagnosis severity groups 1, 2, and 3, respectively. For these cases, the AIDS-defining diagnosis groups can be used to discriminate those with a worse prognosis at diagnosis.

The 80 secondary complications of AIDS were classified into three severity levels, with cut-points determined by clinical judgment. Examples of these complications are shown in Table 2. For each of the three defining diagnosis groups, four subgroups were created, based on the severity of complications (level 0 signified that none of the 80 complications were among the diagnoses on the patient's clinical history). Thus, the matrix includes 12 possible combinations (Table 3).

To estimate the effect of secondary complications on expected survival times, Kaplan-Meier survival curves for each of these 12 combinations were generated and analyzed to construct a hierarchy of combinations of AIDS-defining diagnoses and secondary complications. The resulting index includes four categories (A, B, C, D), with "A" least severe and "D" most severe. Category A had a median survival time of 33 months, whereas severity categories B, C, and D had median survival times of 21, 17, and 12 months, respectively.

Development of a risk group measure. The severity index does not include demographic factors, so the independent effect of these characteristics on survival can be determined. For this project, drug users were identified through the appearance of drug abuse codes and claims for methadone maintenance. Homosexual/bisexual men could not be directly identified, so we grouped all men who did not have evidence of drug use on claims as nondrug users. Similarly, women without drug use claims, who probably acquired the disease through heterosexual transmission, were termed non-drug users; but these non-drug-user groups probably included drug users who were not revealed through our screen. As discussed below, we found the drug-user and non-drug-user groups to be very different in their survival and resource use patterns.

For the pediatric population (< 13 years old), it is also important to determine how they acquired the disease, as children infected perinatally appear to have poorer outcomes than those acquiring the disease through blood transfusions (Krasinski, Borkowsky, and Holzman, 1989). By eliminating children with hemophilia, we were able to group children who develop AIDS within several years of birth as likely to have had

Table 2. Examples of AIDS complications identified and ordered based upon expert judgment

Severity level	Examples
Level 1	Candida esophagitis
	Bacterial pneumonia
	Pulmonary tuberculosis
Level 2	Cytomegalovirus retinitis
	Cryptococcal meningitis
	Toxoplasmosis
Level 3	AIDS dementia complex
	HIV-related cardiomyopathy
	AIDS-associated lymphoma

Table 3. Categories of the severity index for AIDS patients

AIDS-defining diagnosis groups	Subsequent complication severity levels			
	Level 0	Level 1	Level 2	Level 3
Group 1	A	C	D	D
Group 2	B	C	D	D
Group 3	C	C	D	D

Key:AIDS-defining diagnosis groups:
Group 1 includes: Kaposi's sarcoma without organ involvement; mucocutaneous herpes simplex infection: disseminated tuberculosis; candida esophagitis; wasting syndrome and *Salmonella* sepsis.

Group 2 includes: *Pneumocystis carinii* pneumonia.

Group 3 includes: Disseminated atypical mycobacterial infection; Kaposi's sarcoma with organ involvement; cryptococcosis; herpes simplex with organ involvement; disseminated cytomegalovirus infection; cryptosporidiosis/isosporiasis; toxoplasmosis; disseminated histoplasmosis; AIDS-related lymphoma; AIDS dementia complex; disseminated coccidioidomycosis; progressive multifocal leukoencephalopathy.

perinatal transmission of the infection. Yet recent data suggest that some perinatally infected children may live 5 years or more before developing AIDS (Oxtoby, Byers, Simond, and others, 1991). Although we can identify older children assigned ICD-9-CM codes for drug use, we cannot reliably separate perinatally infected children who are older at onset of symptoms from those acquiring the disease through blood products (or rarely, through child abuse).

Examples of Analyses Using the New York Medicaid AIDS Data

Comparison of survival experience of drug users and non-drug users. Several lines of evidence would suggest that drug users should have a worse prognosis after AIDS diagnosis than non-drug users. First, drug users on methadone were observed to have impaired immune function even when they were not infected with HIV (Klimas, Blaney, Morgan, and others, 1991). Second, the CDC recently described higher mortality for drug users of both sexes than for male non-drug users in the same month that they are reported to have AIDS (Berkelman, 1990). Yet analyses of the AIDS survival experience at Montefiore Medical Center showed that, for the group of patients with PCP as the presenting AIDS diagnosis, intravenous drug users (IVDUs) of both sexes have longer survival times than non-IVDUs (Friedland, Saltzman, Vileno, and others, 1991). Differences in survival experience by risk group and gender for AIDS patients are important to explore because they may suggest variations in the impact of HIV infection on specific types of patients, differences in access to health care, or differing responses to treatment.

We used the New York Medicaid AIDS research file to compare the survival experience of patients identified as drug users in our file, as previously described, with those who had no evidence of drug use. We examined only patients with similar patterns of Medicaid enrollment and at least 1 year's followup after AIDS diagnosis. Consequently, we studied 3,805 patients who met the following criteria:

1. A specific CDC AIDS-defining diagnosis appeared in their files;
2. AIDS was diagnosed within the 4-year period of 1983-1987;
3. Medicaid enrollment had occurred more than 6 months before their first AIDS diagnosis; and
4. They were between the ages of 13 and 60.

We tracked the disease experience of these patients for 3 years before the diagnosis of AIDS through 1988 or their death, whichever came earlier.

Patients were grouped into three severity groups based on the type of their first AIDS-defining diagnosis identified on our data base. Of the 3,805 patients in our population, 67 percent were male and 33 percent were female. Approximately equal numbers were identified as drug users (56 and 55 percent, respectively). Men were slightly older than women (36 and 33 years, respectively). The duration of survival was 17 months for drug users and 12 months for non-drug users (logrank statistic, $p < 0.001$). In analyses by gender/risk groups, however, women of both risk groups had minimally different survival times (18 months for drug users versus 15 months for non-drug users, $p = 0.02$), but male drug users survived longer than non-drug users (17 versus 12 months, $p < 0.001$).

In analyses by defining diagnosis severity group, median survival times were: 28 months for severity group 1 (N = 1400; the most common condition in this group was candida esophagitis), 12 months for severity group 2 (n = 1,827; PCP alone), and 7 months for the most severe group (n = 578; the most common conditions were toxoplasmosis and cryptococcal meningitis). To further explore the differing survival experience of male drug users and non-drug users, we compared survival times by severity group. The only significant difference in survival was observed for the least severe group (group 1), 32 versus 18 months for male drug users and non-drug users ($p < 0.001$). We observed that male drug users in the least severe defining diagnosis group were younger at diagnosis than male non-drug users (75.6 versus 72.2 percent aged 13-39 years respectively, $p = 0.004$). In a proportional hazards survival model including clinical and demographic characteristics for the entire population, being younger, a drug user, and in a less severe defining diagnosis group were found to confer a better prognosis.

In summary, these analyses revealed that male drug users had better survival rates for the least severe defining diagnosis group than non-drug users, but age differences appear to contribute to these differences. Comparisons of survival experience by risk group should also consider defining diagnosis, severity, and age.

Patterns of ambulatory care for persons with AIDS. A comparison of the survival experience of the AIDS patients in our data base is only one of the goals of our project. As mentioned earlier, the primary objective is to compare outcomes of specific patterns of care for AIDS. In our preliminary work on this topic, we have observed interesting patterns in the ambulatory care received by our population. For these analyses, we examined all ambulatory visits for our population within a 6-moth period after the diagnosis of AIDS. Our analyses defined three types of ambulatory care patterns: first, visits only to clinics; second, visits only to private physicians; and third, visits both to private physicians and clinics. Only persons with continuous enrollment in the first 6 months after AIDS diagnosis within the period 1983-1988 were analyzed (n= 7,207 AIDS patients). Of this group, 34 percent (2,479) had visits only to clinics, 52 percent (3,779) had visits to both clinics and private physicians' offices, and only 13 percent (949) saw only private physicians within this 6-month period. We found that patients with visits only to clinics tended to be seen in only one clinic, while patients seeing physicians tended to have visits to multiple providers. Only 22.9 percent of the clinic group had encounters in more than one clinic during the period, while 55.2 percent of the physician group saw more than one physician.

The demographic characteristics of the persons with AIDS followed in these settings were found to vary (Table 4). Interestingly, we found that proportionately more children were followed by physicians alone or were seen by physicians and in clinics. In comparing adult populations, low proportions of all risk groups were followed only in private physicians' offices. The "other male" group, who are likely to be homosexual/bisexual men from more affluent backgrounds, were seen only in clinics more than the other groups. It is possible that these clinics specialize in HIV/AIDS care and would be sought by AIDS patients who can afford to "shop around" for care.

As seen in Table 5, women were disproportionately represented in the group getting care from both private physicians and clinics (50 percent for women versus 60 percent for men). Table 5 also shows the proportions of patients in each of three AIDS-defining diagnosis severity groups. Approximately one-third of the group had only an AIDS code (such as 042.x) and no specific AIDS-defining diagnosis. For those who did have an AIDS-defining diagnosis, we examined the distribution among ambulatory care sites. Compared with severity groups 2 and 3, a much larger proportion of those in group 1 (the least severe group) received care both from clinics and private physicians' offices. Interestingly, as the severity of AIDS increases, so do the proportions that are followed in clinics and, to a lesser extent, in private physicians' offices. These patterns suggest that, as

Table 4. Patterns of care for first 6 months on Medicaid with AIDS

Patient characteristic	Clinic only		Physician only		Mixed clinic/physician		Total	
	number	percent	number	percent	number	percent	number	percent
Total group	2,479	34.4	949	13.2	3,779	52.4	7,207	100
Drug user (male)	715	33.9	259	12.3	1,135	53.8	2,109	100
Drug user (female)	225	24.5	111	12.1	581	63.4	917	100
Other males	1,111	41.6	345	12.9	1,214	45.5	2,670	100
Other females	223	27.2	114	13.9	484	58.9	821	100
Children	179	31.3	104	18.2	289	50.5	572	100
Missing	26	22.0	16	13.6	76	64.4	118	100

Note: Includes persons with at least 6 months of enrollment.

the severity of illness increases, patients become increasingly linked with one dominant site of care and tend to have less fragmentation of care across multiple clinics and private physicians' offices. This observation is confirmed when we compare the distribution of care for patients in the first 6 months after AIDS diagnosis with that in the last 6 months of life. In the first 6 months, 34 percent of all patients were treated only in clinics; in the last 6 months 38 percent were cared for only in clinics. The proportion of patients seen only by private physicians is virtually unchanged between the first and last 6 months after AIDS diagnosis (13 and 12 percent, respectively). The shift to care only in clinics suggests that patients are seeking care at sites that are likely to have more broad-based expertise and resources available for the management of complex AIDS-related complications.

Of the 7,207 patients, 1,854 (26 percent) had the majority of their encounters in the first 6 months after AIDS diagnosis with private physicians. Of these visits, 46 percent were to primary care physicians (that is, general internal medicine, family or general practice, and pediatrics), 12 percent were to physicians in specialties that may focus on treating AIDS patients (that is, hematology, oncology, and infectious disease), and 9 percent were to other medical subspecialties. The remainder were to a wide range of physicians including obstetrician-gynecologists, surgeons, and psychiatrists. These data suggest that AIDS patients seeing private physicians are generally followed by primary care physicians.

The majority of patients, however, are followed in clinics only or in conjunction with private physicians. This finding is consistent with a trend seen for all Medicaid patients in States where reimbursement rates for visits are low (Cohen, 1989). New York has historically low reimbursement for physician visits ($11 for an intermediate office visit in 1989) and somewhat higher reimbursement for clinic visits ($15 - $25 in 1989). A special program called the AIDS designated centers instituted in New York in 1986 pays an enhanced rate to specified hospitals to case-manage patients. The Medicaid reimbursement rates for HIV primary care are currently $295 for an initial comprehensive examination and $74 for a routine visit to an AIDS center clinic. It is likely that these AIDS centers treat a substantial portion of Medicaid AIDS patients in their outpatient clinics. Since clinics represent such an important source of care for AIDS, in the future we will explore the impact of specific types of clinics on patient outcomes.

Conclusion

Secondary data bases, such as the Medicaid longitudinal AIDS file that we have described, are invaluable in medical effectiveness research, especially for study of variations and outcomes of health care delivery for persons with chronic diseases. Patients with AIDS or other chronic conditions will tend to receive care from multiple providers over the course of their illnesses. Data collected from only one site may seriously underrepresent the patient's true service use, and efforts to collect these data from multiple providers is time-consuming and technically difficult. A secondary data base that has been longitudinally linked will contain a relatively comprehensive picture of the patient's health care delivery with far less effort. We also believe that the future of research on Medicaid claims data and similar secondary data is promising due to increased attention to the quality of claims information reported.

Of course, secondary data permit analyses of a limited range of predictors of specific types of service use and other outcomes. In our current AIDS project, we are examining the influence of several patient and provider characteristics on patient survival and resource use. We control for the effects of disease severity using a model that incorporates the severity of the first AIDS-defining condition and secondary complications that occur in the first few months after diagnosis. Other important factors, however, that we cannot examine in our models include patient compliance with prescribed treatments, patient satisfaction with health care delivery, and quality-of-life measures. Thus, we expect that our work will

Table 5. Patterns of care for first 6 months on Medicaid with AIDS

Patient characteristic	Clinic only		Physician only		Mixed clinic/physician		Total	
	number	percent	number	percent	number	percent	number	percent
Gender								
Male	1,937	37.6	664	12.9	2,556	49.5	5,157	100
Female	542	26.4	285	13.9	1,223	59.7	2,050	100
Defining diagnosis severity								
Group 1	515	30.0	186	10.9	1,012	59.1	1,713	100
Group 2	921	36.4	303	12.0	1,309	51.6	2,533	100
Group 3	372	39.9	120	12.9	440	47.2	932	100
ARC/AIDS code[1]	671	33.1	340	16.7	1,018	50.2	2,029	100

[1]ARC/AIDS code - patients with AIDS-related complex (ARC) or an AIDS code only and no specific AIDS-defining diagnosis.

Note: Includes persons with at least 6 months of enrollment.

lead to other projects that have access to these primary data.

As discussed in this paper, the challenges and "curses" of working with secondary data bases are substantial. Curses include the inability to confirm suspected cases, censoring, unreliable or missing data, intermittent eligibility, and lack of an applicable severity measure. To overcome these challenges, we have relied on detailed examination of our longitudinal patient histories, linkages with other data sources, use of existing elements in our file as proxies for missing data, and development of new measures applicable to secondary data bases on AIDS. We have not been able to eliminate some problems with our data and our analyses, but we view our work as contributing to a body of information on AIDS care to be supplemented by other researchers.

Medical effectiveness research in HIV disease is in its infancy. In the future, important research initiatives will include study of the role of specific health care technologies for infected patients and the cost-effectiveness of certain treatments. Patients with HIV are living longer, and health care expenditures for the disease are rising (Hellinger, 1991), making it essential that health care resources be used wisely and effectively to produce the best possible clinical outcome and quality of life. We believe this is a sorely needed and rewarding research endeavor that requires the collaboration of teams who can bring many areas of expertise to a project.

Our team comprises clinicians of various specialties, health services researchers, experts in secondary data base research, a biostatistician, medical records experts, computer programming specialists, and Medicaid officials. Similar to the PORT projects supported by AHCPR, the multidisciplinary team enhances the likelihood that sufficient depth of expertise will be available to find and address the curses of these data and to achieve the desired research goal. It is not necessary that all this expertise be housed under one roof. We have found that modern communications make multi-institutional collaboration easy and practical; however, research teams should not function in isolation. Thus, we support the concept of sharing research tools both in the literature and through meetings such as those sponsored by AHCPR that provide a forum for discussion of challenges and solutions in medical effectiveness research. By pooling approaches to deal with the curses of these data, we increase our ability to obtain meaningful results.

References

Andrulis, D.P., Weslowski, V.B., and Gage, L.F. (1989). The 1987 U.S. hospital AIDS survey. *Journal of the American Medical Association*, 262, 784-794.

Ball, J.K., Farley, D.E., and Turner, B.J. (1990). *Does familiarity breed effectiveness in the treatment of AIDS patients? Evidence from a national sample of U.S. hospitals.* Presented at the 118th Annual Meeting of the American Public Health Association, New York, October 1991.

Bennett, C.L., Garfinkle, J., Greenfield, S., and others. (1989). The relation between hospital experience and in-hospital mortality for patients with AIDS-related PCP. *Journal of the American Medical Association*, 261, 2975-2979.

Berkelman, R.L. (1990). Basic epidemiological data on women with HIV. In *Proceedings of the National Conference on Women and HIV Infection.* Bethesda, MD: National Institute of Allergy and Infectious Diseases.

Blumberg, M.S. (1986). Risk adjusting health care outcomes: A methodologic review. *Medical Care Review*, 43, 351-393.

Bravemen, P., Oliva, G., Grisham-Miller, M., and others. (1989). Adverse outcomes and lack of health insurance among newborns in an eight-county area of California, 1982 to 1986. *New England Journal of Medicine*, 321, 513-518.

Carey, T., Weis, K., and Homer, C. (1990). Prepaid versus traditional Medicaid plans: Effects on preventive health care. *Journal of Clinical Epidemiology*, 43(11), 1213-1220.

Carr, E., Coppola, D., and Barry, M.A. (1991). Completeness of outpatient AIDS case reporting - Boston. In *Proceedings of the VIIth International Conference on AIDS* (Abstr. M.C.3269), Florence, Italy. Rome: Instituto Superiore di Sanita.

Centers for Disease Control. (1987). Classification system for human T-lymphotropic virus type III/lymphadenopathy-associated virus infections. *Morbidity and Mortality Weekly Report*, 36, 1s-5s.

Centers for Disease Control. (1991). Update: Acquired immunodeficiency syndrome - United States, 1981-1990. *Morbidity and Mortality Weekly Report*, 40, 358-369.

Centers for Disease Control. (1992). The second 100,000 cases of the acquired immunodeficiency syndrome - United States, June 1981-December 1991. *Morbidity and Mortality Weekly Report*, 41, 28-30.

Cohen, J.W. (1989). Medicaid policy and the substitution of hospital outpatient care for physician care. *Health Services Research*, 24(1), 33-66.

Conviser, R., Grant, C.M., and Coye, M.J. (1991). Update on pediatric AIDS. *Pediatrics*, 87, 642-653.

Demlo, L.K., Campbell, P.M., and Brown, S.S. (1978). Reliability of information abstracted from patients' medical records. *Medical Care*, 16, 995-1005.

Ellwood, M.R., Fanning, T.R., and Dodds, S. (1991). Medicaid eligibility patterns for persons with AIDS in California and New York, 1983-1987. *Journal of the Acquired Immunodeficiency Syndromes*, 4, 1036-1045.

Fanning, T.R., Cosler, L.E., Gallagher, P., and others. (1991). The epidemiology of AIDS in the New York and California Medicaid programs. *Journal of the Acquired Immunodeficiency Syndromes*, 4, 1025-1035.

Friedland, G.H. and Klein, R.S. (1987). Transmission of the human immunodeficiency virus. *New England Journal of Medicine*, 317(18), 1125-1135.

Friedland, G.H., Saltzman, B., Vileno, J., and others. (1991). Survival differences in patients with AIDS. *Journal of the Acquired Immunodeficiency Syndromes*, 4, 144-153.

Gallop, R., Taerk, G., Lancee, W.J., and others. (1989). Knowledge, attitude, and concerns of hospital staff about AIDS. In *Proceedings of the Vth International Conference on AIDS* (Abstr. M.D.P.58), Montreal, Canada. Ottawa: International Development Research Center.

Gerstman, B.B., Lundin, F.E., Stadel, B.V., and Faisch, G.A. (1990). A method of pharmacoepidemiologic analysis that uses computerized Medicaid. *Journal of Clinical Epidemiology*, 43(12), 1387-1393.

Green, J. and Arno, P.S. (1990). The "Medicaidization" of AIDS: Trends in the financing of HIV-related medical care. *Journal of the American Medical Association*, 264(10), 1261-1266.

Hall, T. L., Knickman, J. R., Cross-Dunham, N., and Wittenberg, E. (1989). *AIDS protocol 6: HIV-related training needs, programs, and costs* (NCHSR Program Note). Rockville, MD: Public Health Service.

Hegarty, J.D., Abrams, E.J., Hutchinson, V.E., and others. (1988). The medical costs of human immunodeficiency virus-infected children in Harlem. *Journal of the American Medical Association*, 260, 1901-1905.

Hellinger, F.J. (1990). Updated forecasts of the costs of medical care for persons with AIDS, 1989-93. *Public Health Reports*, 105, 1-12.

Hellinger, F.J. (1991). Forecasting the medical care costs of the HIV epidemic in the United States: 1991-1994. In *Proceedings of the VIIth International Conference on AIDS* (Abstr. W.D.7), Florence, Italy. Rome: Instituto Superiore di Sanita.

Hidalgo, J. (1990). Development of a model longitudinal data base to measure outcomes and quality of care among persons with AIDS. *Quality Review Bulletin*, 16(10), 355-364.

Hohlen, M.M., Manheim, L.M., Fleming, G.V., and others. (1990). Access to office-based physicians under capitation reimbursement and Medicaid case management. *Medical Care*, 28(1), 59-68.

Justice, A., Feinstein, A.R., and Wells, C. (1989). A new prognostic staging system for the acquired immunodeficiency syndrome. *New England Journal of Medicine*, 320, 1388-1393.

Kelen, G.D., DiGiovanna, T., Bisson, L., and others. (1989). Human immunodeficiency virus infection in emergency department patients. *Journal of the American Medical Association*, 262(4), 516-522.

Kelly, J.V., Ball, J.K., and Turner, B.J. (1989). Duration and costs of AIDS hospitalizations in New York: Variations by patient severity of illness and hospital type. *Medical Care*, 27, 423-437.

Kemper, K. and Forsyth, B. (1988). Medically unnecessary hospital use in children seropositive for human immunodeficiency virus. *Journal of the American Medical Association*, 260, 1906-1909.

Keyes, M., Andrews, R., and Mason, M.L. (1991). A methodology for building an AIDS research file using Medicaid claims data and administrative data bases. *Journal of the Acquired Immunodeficiency Syndromes*, 4, 1015-1024.

Klimas, N.G., Blaney, N.T., Morgan, R.O., and others. (1991). Immune function and anti-HTLV-I/II status in anti-HIV-1-negative intravenous drug users receiving methadone. *American Journal of Medicine*, 90, 163-170.

Krasinski, K., Borkowsky, W., and Holzman, R.S. (1989). Prognosis of human immunodeficiency virus infection in children and adolescents. *Pediatric Infectious Disease Journal*, 8, 216-220.

Lemp, G.F., Hirozawa, A.M., Araneta, M.R., and others. (1991). Improved survival for persons with AIDS in San Francisco. In *Proceedings of the VIIth International Conference on AIDS* (Abstr. TU.C.41), Florence, Italy. Rome: Instituto Superiore di Sanita.

Lindan, C.D., Rutherford, G.W., Payne, S., and others. (1989). Decline in rate of new AIDS cases among homosexual and bisexual men in San Francisco. In *Proceedings of the Vth International Conference on AIDS* (Abstr. W.A.P.24), Montreal, Canada. Ottawa: International Development Research Center.

Lurie, N., Ward, N.B., Shapiro, M.F., and others. (1986). Termination of Medi-Cal benefits: A follow-up one year later. *New England Journal of Medicine*, 314, 1266-1268.

Mitchell, J.B. (1983). Medicaid participation by medical and surgical specialists. *Medical Care*, 21, 929-938.

Mitchell, J.B. and Schurman, R. (1984). Access to private obstetrics/gynecology services under Medicaid. *Medical Care*, 22(11), 1026-1037.

Moore, R.D., Hidalgo, J., Sugland, B.W., and Chaisson, R.E. (1991). Zidovudine and the natural history of the acquired immunodeficiency syndrome. *New England Journal of Medicine*, 324(20), 1412-1416.

New York State Department of Health, Division of Epidemiology. (1990). AIDS morbidity and mortality. *New York State Journal of Medicine*, 90, 572-575.

Oxtoby, M., Byers, R., Simond, R.J., and others. (1991). Age at AIDS diagnosis for perinatally-infected children, United States. In *Proceedings of the VIIth International Conference on AIDS* (Abstr. W.C.36), Florence, Italy. Rome: Instituto Superiore di Sanita.

Perloff, J.D., Kletke, P.R., and Neckerman, K.M. (1986). Recent trends in pediatrician participation in Medicaid. *Medical Care*, 24, 749-760.

Phillips, A.N., Lee, C., Elford, J., and others. (1991). P24 antigenemia, CD4 lymphocyte counts and AIDS. In *Proceedings of the VIIth International Conference on AIDS* (Abstr. TU.C.39), Florence, Italy. Rome: Instituto Superiore di Sanita.

Piper, J.M. and Ray, W.A. (1990). Effects of Medicaid eligibility expansion on prenatal care and pregnancy outcome in Tennessee. *Journal of the American Medical Association,* 264(17), 2219-2223.

Raskin, I.E. and Maklan, C.W. (1991). Medical effectiveness research: A view from inside the Agency for Health Care Policy and Research. *Evaluation and Health Professions*, 14(2), 187-200.

Roper, W.L. (1987). AIDS and the role of HCFA. *Journal of the American Medical Association*, 258, 3489.

Rosenblum, L.S., Buehler, J.W., Morgan, M.W., and the Centers for Disease Control Cooperative Surveillance Group. (1991). Completeness of the AIDS reporting in the United States, 1988. In *Proceedings of the VIIth International Conference on AIDS* (Abstr. M.C.3314), Florence, Italy. Rome: Instituto Superiore di Sanita.

Sande, M.A. and Volberding, P.A. (1988). *The medical management of AIDS*. Philadelphia: W.B. Saunders.

Schramm, W.F. (1986). Prenatal participation in WIC related to Medicaid costs for Missouri newborns: 1982 update. *Public Health Reports*, 101(6), 607-615.

Scitovsky, A.A. and Rice, D.P. (1987). Estimates of the direct and indirect costs of acquired immunodeficiency syndrome in the United States, 1985, 1986, and 1991. *Public Health Reports*, 102(1), 5-18.

Smith, M.U., Godeau, R.E., and Katner, H.P. (1989). Knowledge and attitudes of rural Georgia physicians concerning AIDS. In *Proceedings of the Vth International Conference on AIDS* (Abstr. M.D.P.59), Montreal, Canada. Ottawa: International Development Research Center.

Solomon, D.J., Hogan, A.J., Bouknight, R.R., and Solomon, C.T. (1989). Analysis of Michigan Medicaid costs to treat HIV infection. *Public Health Reports*, 104(5), 416-424.

Sullivan, L.W. (1989). The health care priorities of the Bush Administration. *New England Journal of Medicine*, 321, 125-128.

Thaul, S. and Ball, J.K. (1991). *Pediatric HIV-related discharges in a sample of U.S. hospitals: Demographics, diagnoses, and resource use* (AHCPR Pub. No. 91-0031). Rockville, MD: Agency for Health Care Policy and Research.

Thurn, J.R., Campbell, S., and Henry, K. (1989). The impact of AIDS on acute care hospitals: A survey of infectious disease teaching hospitals in the United States. In *Proceedings of the Vth International Conference on AIDS* (Abstr. MDO12), Montreal, Canada. Ottawa: International Development Research Center.

Turner, B.J., Kelly, J.V., and Ball, J.K. (1989). A severity classification system for AIDS hospitalizations. *Medical Care*, 27, 423-437.

Turner, B.J., Markson, L.E., McKee, L.M., and Fanning, T. (1991). Survival patterns of women and men with AIDS: Impact of health care use prior to AIDS. In *Proceedings of the VIIth International Conference on AIDS* (Abstr. TU.D.112), Florence, Italy. Rome: Instituto Superiore di Sanita.

Turner, B.J., Markson, L.E., McKee, L., and others. (1991). The AIDS-defining diagnosis and subsequent complications: A survival-based severity index. *Journal of the Acquired Immunodeficiency Syndromes*, 4, 1059-1071.

Vlahov, D., Munoz, A., Solomon, L., and others. (1991). Prognostic indicators for developing AIDS within 24 months in a cohort of intravenous drug users: Preliminary results. In *Proceedings of the VIIth International Conference on AIDS* (Abstr. TU.C.40), Florence, Italy. Rome: Instituto Superiore di Sanita.

Volberding, P.A., Lagokos, S.W., Koch, M.A., and others. (1990). Zidovudine in asymptomatic human immunodeficiency virus infection. A controlled trial in persons with fewer than 500 CD4-positive cells per cubic millimeter. *New England Journal of Medicine*, 322, 941-949.

Winkelstein, W., Wiley, J., Padian, N., and others. (1988). The San Francisco Men's Health Study: Behaviors associated with incident human immunodeficiency virus infection among homosexual/bisexual men. *American Journal of Public Health*, 78, 1472-1474.

WITS Study Group. (1991). Characteristics of a cohort of HIV-seropositive pregnant women recruited into the study of women and infant transmission. In *Proceedings of the VIIth International Conference on AIDS* (Abstr. WC-3256), Florence, Italy. Rome: Instituto Superiore di Sanita.

Zaccharelli, M., Pezzotti, P., Salmaso, S., and others. (1991). Evaluation of the impact of the 1987 AIDS case definition. In *Proceedings of the VIIth International Conference on AIDS* (Abstr. M.C.3167), Florence, Italy. Rome: Instituto Superiore di Sanita.

Overcoming the Curses of Imperfect Data

Fred J. Hellinger, Ph.D.

Introduction

The paper by Turner, Markson, Houchens, and Fanning (in this volume) discusses the use of a 6-year longitudinal file created by the New York State Department of Social Services to examine the provision of care to persons with acquired immunodeficiency syndrome (AIDS) (PWAs). This file includes data from nearly 8 million Medicaid claims and eligibility records from 1983 to 1988 and contains information on about 16,000 PWAs. Currently, Turner is conducting a study to determine, using this data set, the effects of continuity of care on the cost of care and progression of illness for homosexual men and intravenous drug-using men, women, and children.

Accuracy of Diagnostic Codes

One of the major limitations of using this data set to examine the continuity of care of PWAs is the accuracy of diagnostic codes. Accuracy has been a particular problem in assigning diagnostic codes for AIDS because of the complexity of the Centers for Disease Control (CDC) definition of AIDS and because of changes in the available codes and in the definition of AIDS. A fundamental concern is the ability of researchers to identify PWAs using diagnostic codes from claims records.

In a study of 6 million hospital discharge records from 21.5 percent of short-term general hospitals in the United States in 1987, 11,477 had an AIDS-related code as either a principal or secondary diagnosis (Commission on Professional and Hospital Activities, 1988). This study found that 3,628 records included the code for pneumocystosis (136.3), but more than 20 percent of these records included no other AIDS-related code. Numerous other coding inconsistencies were found, and the study concluded that, "the data collection for this AIDS Supplement indicates that there is considerable variation in the manner in which AIDS is reported" (p. 17).

A study conducted by the California Department of Health Services found that about 10 percent of persons with AIDS who were identified from the State's AIDS registry were hospitalized and that there were no AIDS-related codes on the discharge records (Hiehle, Maxfield, and Kizer, 1990). These persons were referred to as "clinically veiled cases." Hidalgo (1990) examined the accuracy of International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) codes in Maryland and concluded that, "reliance upon ICD-9-CM coding in hospital discharge abstracts to identify PWAs may result in unreliable findings unless confirmation is obtained through other information sources" (p. 15). Keyes, Andrews, and Mason (1991), stated that:

> Prior to 1986, there were no unique ICD-9-CM diagnosis codes assigned to AIDS. Although unique codes for AIDS were developed and implemented in October 1986, there was speculation that such codes were often not used correctly in routinely reporting systems such as hospital records and claims. (p. 3)

Indeed, much of the confusion with the use of the 42-44 codes may be attributed to the complexity of the instructions for their use. The publication (Centers for Disease Control, 1991) that describes these codes states:

> Only one code for the 042-044 series should be used. For instance, a patient with candidiasis of the lung (112.4), Kaposi's sarcoma (176.0-176.9), and HIV infection described as AIDS would be assigned only one 042 code ... If a person has ever been diagnosed as having AIDS or any of the HIV manifestations classifiable to 042, a code from 042 should be assigned for all subsequent admissions even if the patient is currently under treatment for a condition that would otherwise be assignable to 043, or a code from 044. Similarly, if a patient has ever been diagnosed with ARC or any condition classifiable to 043, a code from 044 should not be used. In these in-

Dr. Hellinger is Director, Division of Cost and Financing, Center for General Health Services Extramural Research, Agency for Health Care Policy and Research, Rockville, MD.

> stances, a fourth digit of .9 ("unspecified") should be used with the HIV infection code. (p. 3)

It is improbable that hospital coders rigorously adhere to these instructions. Indeed, a coder would need access to all medical records (hospital and ambulatory) in order to abide by these instructions. This system is unrealistic, and coding errors for patients with human immunodeficiency virus (HIV) are common.

The ideal method of dealing with coding errors is to use alternate sources of information (such as the AIDS Registry and death certificates) to help identify PWAs. Researchers in California and Michigan used AIDS registry files and death registry data to identify PWAs in their Medicaid programs. In addition, researchers in Michigan used payment records with pharmacy payment for zidovudine to help identify PWAs (Solomon, Merwin, and Hogan, 1990).

Turner and coworkers should examine the possibility of using New York State's Medicaid rate codes identifying services provided to AIDS patients or claims with the formulary code for zidovudine capsules to help identify persons with HIV/AIDS or to confirm the validity of diagnostic codes. For example, a person who was hospitalized with a diagnostic code of 783.2 (abnormal weight loss) who also received zidovudine capsules might be assumed to be a PWA.

Turner and coworkers should discuss how inclusion of the codes for AIDS/HIV (42-44) described in the October 1986 addendum to the ICD-9-CM classification system affected the accuracy of coding AIDS and HIV (without AIDS) patients. In addition, they should address the issue of whether separate screening algorithms need to be developed for patients seen before and after October 1986. They also should discuss how the change in the AIDS definition in 1987 affected the accuracy of AIDS/HIV codes. After September 1987, patients diagnosed presumptively with *Pneumocystis carinii* pneumonia (PCP), cytomegalovirus retinitis, Kaposi's sarcoma, and several other conditions were included as AIDS patients, whereas before September 1987, they were not.

Turner and colleagues also should address problems in accuracy of coding for the HIV-related codes 42-44. There is evidence that PWAs were coded as 43 and 44, especially during the period immediately after October 1986. The Commission on Professional and Hospital Activities (1988) length of stay study stated, "With AIDS patients presenting many disease manifestations, confusion may exist over the selection of the appropriate code...despite the fact that the October 1986 Addendum to the ICD-9-CM classification states that codes in the categories 042, 043, and 044 are mutually exclusive, there is evidence of multiple coding of AIDS diagnoses" (p. 17).

Solomon, Merwin, and Hogan (1990), in a study of HIV/AIDS using Michigan Medicaid payment records, found that many persons who received an ELISA test were coded for an HIV-related illness. Their statement that, "It is common practice for providers to record the diagnosis which they are looking for even though the results are negative due to the belief it increases the likelihood of payment" (p. 2) demonstrates the influence of payment on coding accuracy.

The AIDS Institute of the New York State Department of Health stated that, "A 5-tier enhanced reimbursement schedule was instituted in 1989 to encourage earlier diagnosis and improve access to HIV counseling and testing, clinical evaluation, and PCP prophylaxis in the first stage of HIV infection...In mid 1991, the primary care rates were extended to individual practitioners in the community" (1991). Since AIDS services are reimbursed at an enhanced rate by Medicaid in New York State, Turner and coworkers should discuss the potential effects of the enhanced rates on the accuracy of HIV-related codes.

In addition, it may be that diagnostic codes reported on outpatient claims are less accurate than diagnostic codes reported on inpatient claims. This is because diagnostic codes on inpatient claims directly determine reimbursement (diagnosis-related groups are defined using inpatient diagnostic codes) and diagnostic codes on outpatient claims do not directly determine reimbursement (outpatient care is reimbursed on the basis of the type of service provided, not on the patient's diagnostic codes).

Turner and colleagues (in this volume) use the case-finding method described by Keyes, Andrews, and Mason (1991) to create a list of ICD-9-CM codes to identify suspected AIDS cases in the Medicaid population. A diagnostic algorithm is derived based on a comparison of the ratio of verified (that is, cases confirmed using the California AIDS Registry data) to nonverified cases. In other words, the diagnostic criteria used to identify AIDS cases are derived by selecting the set of diagnostic codes most likely to be associated with verified AIDS cases.

The algorithm derived by Keyes and colleagues (1991) yielded 80 percent verified cases for disabled males age 18-50, 60 percent for nondisabled males age 18-50, 65 percent for females age 18-50, and only 50 percent for children. Thus, of the children who fit the diagnostic criteria, only about one-half were confirmed using the California AIDS registry, and more than one-third of the women were not confirmed. This is not surprising because, of the 1,152 persons in the Medi-Cal population meeting the diagnostic criteria in 1985, only

10 were children and only 45 were females. Hence, the diagnostic criteria were based almost entirely on the experience of the 1,034 disabled men age 18-50.

Turner and colleagues (in this volume) state that, "Our case-finding protocol appears to function well primarily because of being able to test it with registry data in California." It is unclear whether the algorithm derived using the codes assigned to Medi-Cal patients will work as well for Medicaid patients in New York State. It also is unclear how well (or poorly) the diagnostic criteria will work for women and children in the New York State Medicaid program. This issue is important for the AIDS population in New York State, which comprises a higher proportion of women and children.

Severity Index

Turner and colleagues (in this volume) use a complex method to derive a severity of illness index for adults with AIDS (SIAA) which can be used with longitudinal data. The SIAA uses a combination of expert clinical judgment and statistical techniques to identify sets of symptoms and manifestations that exhibit similar survival times. The final configuration of the SIAA uses the AIDS-defining diagnosis to place a patient in one of three severity groups and any subsequent complications to place the patient in one of four severity-of-complication groups. The resulting 12 subgroups are examined in terms of survival times and placed into one of the final four survival groups (A, B, C, or D) (see Table 3, Turner, Markson, Houchens, and Fanning, in this volume). The mean survival times for groups A, B, C, and D are 33, 21, 17, and 12 months, respectively.

The mean survival time for persons with Kaposi's sarcoma and "all other" manifestations is 14 months, and the mean survival time is 10 months for persons with PCP. It is clear that the SIAA helps specify cases with diverse survival times within these two groups; however, the SIAA defines cases using an algorithm that combines very different types of conditions. The severity of AIDS is not easily defined on a linear scale from least to most severe, and most classification systems appear artificial.

Turner and coworkers use the SIAA to examine the survival experience of drug users and nondrug users. Male drug users and male nondrug users survived about the same period of time; however, in the least severe diagnosis group, drug users survived a longer period of time.

This analysis suffers from a right censorship problem. An unknown number of individuals leave New York State, and some of these individuals die. Access to National Death Index data would be useful in mitigating this problem (Turner and coworkers acknowledge this problem and the importance of the National Death Index to help track patients). It also might be useful to examine the survival times of individuals who cease to have claims before the end of the study period but who have not been identified as deceased. These individuals may have left New York State, and their inclusion in this study may lead to overestimates of survival times.

A major limitation in the use of Medicaid data to analyze persons with HIV is that many individuals are eligible during only part of the year. Turner and colleagues (in this volume) provide a good discussion of this problem and note that only 60 percent of the adults with AIDS on Medicaid were continuously enrolled for the first 6 months after enrollment. A recent publication (New York State Department of Health, 1990) reveals that the average PWA on Medicaid was eligible 9.2 months during fiscal year 1988. Turner and coworkers found that continuity of care was high during the last 6 months of life for all groups except male nondrug users. This finding is not surprising because nondrug users are more likely than drug users to exceed the income or asset requirements of Medicaid.

Conclusion

It is evident that the use of Medicaid data to examine the experience of PWAs entails many problems. Medicaid, however, is one of the most extensive sources of information about PWAs and one of the best sources of data about women and children with HIV. Turner and colleagues (in this volume) acknowledge many of the problems of working with Medicaid data and adjust their analyses appropriately, when possible, to compensate for the deficiencies of this data source. Because primary data collection for those with AIDS is costly, most important questions about the care of people with AIDS (for example: What types of services are received by PWAs? How does the care received by males and females differ? How does the care received by drug users and non-drug users differ?) are addressed using secondary data sources. Other questions such as the impact of insurance on the type of care received may only be addressed using data sets, such as State discharge data sets, obtained from individuals with different types of coverage.

Given the likely adoption of a new definition of AIDS that includes anyone who is HIV positive and has 200 or fewer T-helper cells, it is not too early to examine the impact of this possible change on studies of PWAs. In particular, such a change would enable large numbers of people to qualify for Medicaid who heretofore were in-

eligible for Medicaid because they were not designated as disabled.[1]

References

AIDS Institute of the New York State Department of Health. (1991, June). *New York State HIV-Medicaid rates.* Albany, NY: Department of Health.

Centers for Disease Control. (1991, July 26). Human immunodeficiency virus (HIV) infection codes and new codes for Kaposi's sarcoma: Official authorized addenda ICD-9-CM. *Morbidity and Mortality Weekly Report*, 40(RR-9), 3.

Commission on Professional and Hospital Activities. (1988). *Length of stay: AIDS Supplement 1987.* Ann Arbor, MI: Author.

[1]A diagnosis of AIDS is currently a presumptive indicator of disability for the Social Security Administration—it is not at all certain that it would continue to be so if the definition of AIDS is changed to include all HIV-positive individuals with fewer than 200 T-helper cells.

Hiehle, G., Maxfield, W.T., and Kizer, K.W. (1990, March). *Medi-Cal studies in AIDS demographics and expenditures for persons with AIDS, 1980-1989.* Sacramento, CA: California Department of Health Services.

Keyes, M., Andrews, R., and Mason, M. (1991). A methodology for building an AIDS research file using Medicaid claims and administrative databases. *Journal of Acquired Immune Deficiency Syndromes*, 4(10):1015-1024.

New York State Department of Health. (1990). *AIDS in New York State.* Albany, NY: Department of Health.

Solomon, D., Merwin, S.D., and Hogan, A. (1990). *Analysis of the cost of treatment for HIV infection through the Michigan Medicaid Program.* Paper presented at the meeting of the American Public Health Association, New York, NY.

Turner, B.J., Principal Investigator. (1991). *Consequences of patterns of provider care for AIDS*, interim report on grant HS-06455 to the Agency for Health Care Policy and Research, Rockville, MD.

Getting the Most Out of Messy Data: Problems and Approaches for Dealing with Large Administrative Data Sets

Patrick S. Romano, M.D., M.P.H.
Harold S. Luft, Ph.D.

Introduction

Administrative health care data bases, which have been created "primarily to monitor utilization, to determine the consumption of resources, (or) to ascertain the capacity to supply services" (Agency for Health Care Policy and Research, 1991), provide an increasingly accessible and widely used source of data for health care research. These data sources have become especially important in recent years, as public and private research sponsors have encouraged investigators to study the appropriateness and effectiveness of health services across hospitals and across communities. Administrative data bases are indispensable for such research because they can provide large, representative or universal samples from wide areas at low cost. Unfortunately, the usefulness of administrative data bases is limited by incomplete records, unreliable or invalid coding, and missing clinical variables that may represent important predictors of outcomes.

This paper will identify the major problems that arise in using administrative data to analyze health outcomes and present methods for overcoming these problems. We have chosen to focus on hospital discharge abstracts and Medicare claims data because the most experience has been accumulated with these types of data. Our comments, however, may be equally applicable to data collected by other Government agencies or private insurers. The problems to be discussed include the completeness of the sampling frame from which cases are drawn, the inherent deficiencies of the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) classification, the reliability and validity with which diagnoses and procedures are coded, the difficulty of linking sequential hospitalizations or visits, and the unknown timing of multiple diagnoses listed on a single record.

Dr. Romano is Assistant Professor of Medicine in Residence, University of California, Davis. Dr. Luft is Professor of Health Economics, Institute for Health Policy Studies, University of California, San Francisco.

Problems with the Sampling Frame

Nature of the problem. The major strength of administrative data bases is their inclusiveness. Because hospitals are required to submit data either by statute or as a condition of reimbursement, these data bases are usually quite complete. As a result, an investigator who uses administrative data can generally be confident that his or her sampling frame resembles the universe of patients whose insurance status or location qualifies them for inclusion in that data set. When the administrative data base is missing cases that ought to be included, population-based estimates of disease incidence or health care utilization may be biased, and the sample may be unrepresentative. This problem can arise in several ways.

If the population of elderly Medicare enrollees corresponds closely to the elderly population of the United States, then accurate estimates of hospital utilization by the aged can be obtained from Medicare discharge data alone. Fisher, Baron, Malenka, and others (1990), however, compared Census Bureau estimates of the total U.S. population over age 65 with estimates derived from the Medicare enrollment file (HISKEW) and reported that Medicare enrollees represented approximately 96 percent of the estimated elderly population on July 1, 1985. Thus, the number of persons aged 65-69 years is underestimated, perhaps because of the delayed enrollment of persons who postpone retirement. In fact, the underestimation of the total elderly population from Medicare data may be even greater to the extent that the Census Bureau's mid-decade estimates reflect systematic undercounting in the 1980 decennial census.

This confusion about the size of the sampling frame is particularly problematic with private insurance claims data. Indemnity insurers frequently do not know the number of insured persons in employer-sponsored group plans; they may estimate the number of dependents and allow employers to record additions and terminations. Even those insurers that monitor the size of their enrolled populations often do not track the eligibility status of individuals. Without such information, a researcher cannot generate a longitudinal profile of an individual's utilization and outcome.

Dual insurance coverage represents a potentially important cause of missing cases in claims data bases. For example, a married person in a two-income household may enjoy coverage under two different insurance plans. The reimbursement claim for a specific episode of care may be submitted only to the primary insurer. Similarly, some Medicare and Medicaid beneficiaries receive care in Department of Veterans Affairs (VA) medical centers. Approximately 4 percent of presumed hip fracture cases in men in New England were identified from VA files (Fisher, Baron, Malenka, and others, 1990); these cases would not have been reported to a claims data base unless they were first admitted to a non-Federal hospital. A specific problem with Medicare Provider Analysis and Review (MEDPAR) discharge data is that records of Medicare beneficiaries enrolled in health maintenance organizations (HMOs) are sometimes not submitted. In fiscal year 1990, about 1.2 million Medicare beneficiaries were enrolled in these plans, under which Medicare pays a capitated monthly payment for all necessary medical care. Since these risk-based HMOs pay the entire hospital bills for their enrollees, neither the HMO nor the hospital has an incentive to submit copies to the Medicare fiscal intermediary. Effective August 1988, the Health Care Financing Administration requested that hospitals submit no-pay bills, but the Government Accounting Office (1991) recently found that intermediary data still include only about 61 percent of the discharges suggested by a review of records at 10 large HMOs.

Estimation of the appropriate denominator is particularly difficult with Medicaid data for two reasons. First, most States use Medicaid funds to purchase Part B coverage for their elderly poor. Claims for these double enrollees are paid primarily by Medicare, so the utilization data in Medicaid files may be incomplete. Second, the eligibility status of Medicaid beneficiaries changes from month to month; Medicaid enrollment files may not record these changes in a timely manner (Bright, Avorn, and Everitt, 1989).

Hospital claims or discharge abstracts may be lost in the process of coding or transmission. For example, Fisher, Baron, Malenka, and others (1990) could not locate discharge records for some Medicare enrollees whose Part B claims suggested that they had been admitted with acute hip fracture. They searched the 5-percent national MEDPAR file from 1985 for all records with a discharge diagnosis of hip fracture and the 5-percent Medicare Part B file for claims documenting an acute hip fracture; over 9 percent of all hip fracture cases were identified only from Part B data. This percentage varied from 0 percent in Connecticut to 33 percent in Virginia, suggesting that the policies and performance of fiscal intermediaries or other regional factors may affect the completeness of MEDPAR data.

With changing technologies and increasing emphasis on cost containment in recent years, more procedures are being performed on an outpatient basis than ever before. Gastrointestinal endoscopies, cataract extractions, and coronary catheterizations are now being done routinely in outpatient settings. Sampling frames derived from inpatient data systematically exclude substantial numbers of these procedures. Procedures performed in an emergency room before admission may also be excluded.

Approaches to the problem. In calculating population-based rates of morbidity or health care utilization, investigators should be careful to use a denominator that reflects the actual sampling frame. Hence, the denominator for Medicare-based analyses should generally be the number of Medicare-enrolled persons in an area, rather than the Census population estimate (Fisher, Baron, Malenka, and others, 1991). For studies using comprehensive hospital discharge abstracts, Census estimates are generally preferable (although the presence of military personnel may lead to overestimation of the population at risk for admission to non-Federal hospitals). When the elderly population is of particular interest, Medicare enrollment files offer a reasonable alternative. Medicare-derived estimates of the elderly population are apt to be more reliable at the local level than Census-derived estimates, especially in communities that have experienced substantial demographic change since the last decennial census. In addition, it is easier to determine the number of aged residents in each zip code from Medicare data than from the Census; such estimates may be useful for defining hospital market areas from administrative data. Medicare-derived population estimates are unreliable, however, wherever large concentrations of persons outside the Social Security system (for example, former Government employees or recent immigrants) reside. Unfortunately, both Medicare and Census data are likely to undercount the actual population in areas with many homeless or undocumented persons.

The proper denominator for a study limited to claims data would be the number of persons for whom that insurer is the primary source of coverage, rather than the total number of covered individuals. The availability of such information varies by insurer, although Medicare is not supposed to be the primary payer in cases of overlapping coverage. Whenever possible, individuals with secondary coverage by an insurer should be removed from both the numerator and the denominator in calculating incidence or prevalence for an insured population. This adjustment minimizes the bias introduced by systematic underreporting of claims to the secondary payer.

If multiple data bases can be linked, then missing records can be identified from another data source. Fisher, Baron, Malenka, and others (1990) described a method for validating the frequency of hip fractures by searching both Part A and Part B Medicare claims. Roos, Roos, Cageorge, and Nicol (1982) performed a similar analysis using parallel claims submitted to the Manitoba Health Services Commission by surgeons, anesthetists, and assisting physicians. They found that three identical tariffs were filed in 91 percent of cases involving hysterectomy and in 89 percent of cholecystectomy cases. Even if direct linkage is impossible, the aggregate totals can be compared. For example, the number of cholecystectomies performed on elderly Californians could be estimated both from hospital discharge abstracts compiled by the Office of Statewide Health Planning and Development (OSHPD) and from MEDPAR data.

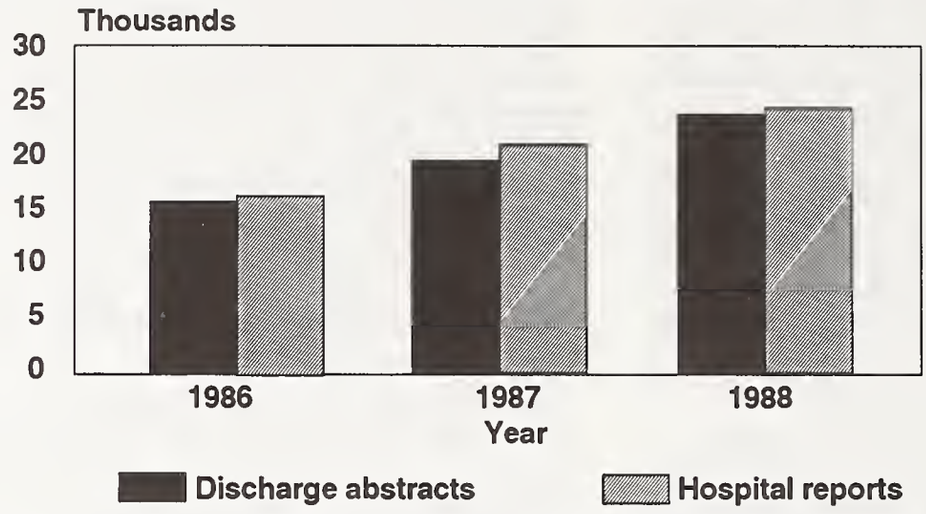
Estimates of procedure or diagnosis frequency derived from secondary data sources can be compared with each hospital's or selected hospitals' tallies. For example, California hospitals specifically report how many open heart surgeries and percutaneous transluminal coronary angioplasties (PTCAs) they do each year (OSHPD, 1989). According to this source, 24,565 PTCAs were performed in California in 1988 (Figure 1). Hospital discharge abstracts revealed 24,190 inpatient occurrences of PTCA (based on a principal or secondary procedure code of 36.01-36.02 or 36.05); an additional 134 likely occurrences were identified by the code 36.00 (unspecified removal of coronary artery obstruction). The 241 missing cases, representing a discrepancy of less than 1 percent, could include procedures performed on outpatients or on inpatients discharged in 1989, as well as procedures that were miscoded on discharge abstracts. In addition, some hospitals may have overestimated their PTCA caseload. Of course, we cannot assume that all cases identified from discharge abstracts were included in the tallies provided to OSHPD; the actual total may exceed either estimate. When we initially prepared Figure 1, we found a larger discrepancy in 1987; closer examination revealed that a major PTCA hospital had been inadvertently excluded because all of its discharges were labeled psychiatric rather than acute care in that year! Trauma centers maintain registries documenting the number of patients meeting accepted criteria as major trauma victims. These figures can be compared with those obtained from claims data, although major trauma is difficult to identify reliably in claims data bases.

Problems Arising from the ICD-9-CM Classification Scheme

Nature of the problem. McMahon and Smits (1986) described three major problems with the ICD-9-CM coding system that has been used since 1978 to characterize diagnoses and procedures in most administrative data sets. First, ICD-9-CM often fails to distinguish clinically dissimilar diagnoses because of inherently imprecise codes. For example, the code for acute myocardial infarction (AMI) includes any myocardial ischemia "specified as acute or with a stated duration of 8 weeks or less." Before this definition was modified in 1989, patients admitted for emergency management of myocardial infarction (MI) received the same code as those admitted for elective evaluation up to 2 months thereafter. Patients with postinfarction angina, a diagnosis that confers a particularly high risk of recurrent infarction, cannot be distinguished from patients with chronic stable angina. ICD-9-CM codes characterize the site of lung cancer but not the histology, despite the greater prognostic significance of the latter variable. The specificity of ICD-9-CM coding varies dramatically by system; it is greater for ocular abnormalities than for more life-threatening gastrointestinal abnormalities. It permits specification of the site of MI but precludes specification of the site of cerebral infarction.

A related source of ambiguity in ICD-9-CM coding is the availability of nonspecific codes as alternatives to more precise codes. This characteristic allows unsophisticated or unmotivated coders to use general codes when more specific codes would be desirable. For example, "unspecified septicemia" may be used in place of the more precise diagnosis, "streptococcal septicemia." Overlooking the cause of an infectious disease was found to be particularly frequent in a recent study of California discharge abstracts, in which a randomly selected sample of records was reabstracted by an expert coder (OSHPD, 1990). Another example of this phenomenon is the frequent use of the code 410.9 (acute myocardial infarction, not otherwise specified). Blumberg (1991) reported that 44 percent of U.S. hospitals coded over 20 percent of their Medicare AMI cases as "not otherwise specified" in 1987.

Figure 1. Comparison of PTCA frequency in California 1986-1988, determined by two methods



Source: Hospital reports were derived from the annual survey form that all California hospitals are required to complete.

Even when precise codes are employed, their clinical meaning is often unclear. As Iezzoni (1990) described, the level of hematocrit that justifies the diagnosis of anemia is not specified. Similarly, vague diagnoses such as respiratory failure and coronary atherosclerosis are not defined. Patients in different regions may therefore receive different diagnosis codes for the same clinical problem. For example, New Jersey hospitals in 1982 frequently applied the diagnosis "angina pectoris," whereas North Carolina hospitals preferred "coronary atherosclerosis" in describing similar patients. Such variation reflects the fact that conditions can be coded according to their pathophysiology (coronary atherosclerosis), clinical manifestation (ischemic heart disease), or symptoms (chest pain) (Steinberg, Whittle, and Anderson, 1990).

A second major problem with the ICD-9-CM coding system is that the surgical approach often cannot be specified (McMahon and Smits, 1986). This deficiency makes it difficult to compare surgical outcomes across hospitals or regions, since all patients did not undergo the same procedure. For example, mesocaval, portocaval, and distal splenorenal shunts are all described as "intra-abdominal venous shunts" (39.1). The code for surgical fracture reduction of "other specified bone" includes repair of pelvic, sternal, rib, and scapular fractures (Mullin, 1985). Until 1989, the same code was used to describe both surgical and endoscopic destruction of intestinal lesions.

Finally, several misleading coding conventions are applied to the ICD-9-CM classification scheme. As Feinstein (1988) noted, "no instructive guidelines have ever been developed as operational criteria for how and when to apply the available nomenclature in diagnosis of individual patients." Perhaps the best example is that patients admitted for "rule out myocardial infarction" may be coded as AMI even if AMI is ruled out by clinical criteria. Iezzoni, Burnside, Sickles, and others (1988) showed that 42 percent of the patients with a discharge code of 410 at five Boston-area teaching hospitals did not meet clinical criteria for AMI, compared with 9 percent at five nonteaching hospitals. The most important reason for this discrepancy was the misclassification of patients who were ruled out for AMI. Kennedy, Stern, and Crawford (1984) also found that such miscoding contributed substantially to the apparent overdiagnosis of AMI in patients at one Texas medical center. Another misleading convention is that hypertensive patients admitted for congestive heart failure are classified under an etiologic heading (hypertensive heart disease), whereas similar patients without hypertension are classified under a clinical syndrome (congestive heart failure).

Approaches to the problem. Searching for associated diagnoses, procedures, or characteristics of the hospital stay may help identify patients who had the specific problem of interest. For example, it would be implausible for someone to be discharged home alive after an AMI in less than 3 days. That scenario is more consistent with a ruled-out MI. The location of a cerebral infarct may be surmised from associated codes for neurological deficits (for example, aphasia or hemiplegia). If the depth of a traumatic liver injury is not specified (ICD-9-CM 864.00), then the associated procedure codes (for example, primary operative repair versus wedge or lobar resection) can be used to distinguish complex from simple lacerations. Conversely, associated diagnosis codes can be used to distinguish identi-

cally coded procedures involving different structures. A patient undergoing open reduction with fixation of "other" bones should have a more precise code for the underlying diagnosis, such as pelvic fracture. Tunis, Bass, and Steinberg (1991) developed a clever approach to identify occurrences of an ambiguously coded procedure (percutaneous transluminal angioplasty of the lower extremities) by searching for diagnoses that represent "likely" or "possible" indications for the procedure (for example, peripheral vascular disease).

Procedures also may be used as surrogates for diagnoses when the diagnoses are apt to be coded imprecisely. For example, preoperative use of an intra-aortic balloon pump can be used as a surrogate for severe ventricular dysfunction. Valve replacement concurrent with coronary artery bypass grafting (CABG) can serve as a surrogate for severe valve disease. Obviously, this strategy works only if the date of a secondary procedure relative to the primary procedure is known. Any procedure used as a surrogate for a diagnosis should have relatively clear indications; hence, carotid endarterectomy would be a poor surrogate for cerebrovascular disease (Winslow, Solomon, Chassin, and others, 1988).

Comorbidities or complications of interest should be defined so as to encompass comparable specific and nonspecific codes. Questions about the aggregation of codes can be answered by examining the number of patients with individual codes and their outcomes. For example, patients in California with unspecified intracranial hemorrhage have similar mortality as those with specified subarachnoid or subdural hemorrhage (Table 1). These diagnoses can be aggregated safely. On the other hand, patients with pelvic fracture of unspecified type (22 percent of all pelvic fractures) have twice the mortality of patients with acetabular, pubic, iliac, ischial, or other specified fractures. Hence, specified and unspecified pelvic fractures cannot be treated as equivalent diagnoses.

Similarly, Jencks, Williams, and Kay (1988) found that mortality following unspecified MI (ICD-9-CM 410.9) substantially exceeded that following specified anterior or subendocardial MI in Medicare discharges in 1985. So should we use unspecified site as a risk factor in modeling outcomes? If the observed difference in risk reflects a true difference in the size or location of the infarct, then ignoring the distinction may introduce an

Table 1. In-hospital mortality rate by anatomic site for pelvic fractures and traumatic intracranial hemorrhages requiring hospitalization in California, 1983-1988

Diagnosis (anatomic site)	ICD-9-CM codes	Mortality rate (percent)	Number of cases
Subarachnoid hemorrhage	852.0, 852.1	22.0	2,309
Subdural hemorrhage	852.2, 852.3	19.9	9,897
Extradural hemorrhage	852.4, 852.5	5.6	2,032
Multiple specified sites	As above	28.2	681
Other and unspecified intracranial hemorrhage, with or without fracture	853, 800.2, 800.3, 800.7, 800.8, 801.2 801.3, 801.7, 801.8, 803.2, 0803.3 803.7, 803.8, 804.2, 804.3, 804.7, 804.8	19.9	16,823
Fracture, acetabulum	808.0, 808.1	2.7	8,758
Fracture, pubis	808.2, 808.3	2.7	18,132
Fracture, ilium	808.41, 808.51	2.3	2,284
Fracture, ischium	808.42, 808.52	2.1	1,648
Fracture, other specified	808.49, 808.59	2.8	1,227
Multiple specified fractures, no disruption of pelvic circle	As above	3.4	4,009
Multiple fractures with disruption of pelvic circle	808.43, 808.53	7.4	3,285
Fracture, unspecified site	808.8, 808.9	4.6	11,245

Source: Derived from discharge abstracts for all patients with a diagnosis of pelvic fracture (ICD-9-CM 808) or intracranial hemorrhage (ICD-9-CM 852, 853, 800.2, 800.3, 800.7, 800.8, 801.2, 801.3, 801.7, 801.8, 803.2, 803.3, 803.7, 803.8, 804.2, 804.3, 804.7, 804.8) in California acute-care non-federal hospitals in 1983 through 1988. All selected patients were assigned to one of the above categories based upon their ICD-9-CM diagnosis codes.

omitted variable bias. Alternatively, unspecified site may be an endogenous variable if it is associated with mortality only because clinicians have difficulty localizing rapidly fatal infarcts (Blumberg and Binns, 1989). Including unspecified site as a predictor in mortality regression models would then lead to overfitting. The correct strategy is not apparent, especially if the nonspecific code is correlated with variables of critical interest, such as hospital characteristics. Blumberg and Binns (1989) excluded patients from hospitals that had extreme proportions of unspecified AMI from their case-mix adjustment models to avoid rewarding these facilities by improperly inflating the predicted risk for their patients. Although such an exclusion seems reasonable, the boundary proportions must be carefully defined, or prognostically important case-mix differences may be overlooked.

Particularly imprecise ICD-9-CM codes should probably be ignored. Our analysis of data from nine California hospitals that had exceptionally high or low risk-adjusted rates of aggregated postoperative complications revealed that the most imprecise ICD-9-CM diagnoses (for example, "cardiac complications during or resulting from a procedure" or 997.1) varied the most in prevalence between low-morbidity outliers and high-morbidity outliers (Table 2). More specific complication codes (for example, "disruption of operation wound") differed little in prevalence between these two groups. Although CABG cases with code 997.1 in New York were 3.4 times more likely to have had substandard care than those without this code (Hannan, Bernard, O'Donnell, and Kilburn, 1989), the extreme interhospital variability illustrated in Table 2 is not plausible. This situation presumably reflects not differing quality of care, but differing coding practices for nonspecific complications. In comparing hospital or regional outcomes, one should focus on complications that are specific and well-defined and have known clinical significance.

Problems with the Reliability and Validity of ICD-9-CM Coding

Nature of the problem. Several types of errors may occur in the process of assigning ICD-9-CM codes to patients' actual diagnoses and procedures (Hsia, Krushat, Fagan, and others, 1988). Misspecification occurs when the attending physician lists inaccurate, incomplete, or unsupported diagnoses. Miscoding occurs when medical records personnel misinterpret the physician's notes and apply incorrect ICD-9-CM codes. This phenomenon includes overlooking diagnoses listed in the chart and coding diagnoses that are not listed. Miscoding also includes keypunch errors such as transposed digits. Missequencing occurs when a secondary diagnosis or procedure is substituted for the correct primary diagnosis or procedure. Missequencing is relevant only

Table 2. Frequencies of ICD-9-CM defined post-operative related complications among coronary bypass graft patients in California in 1986, by hospital outlier status

Diagnosis	ICD-9-CM codes	Five negative outlier hospitals, percent (N=2,384)	Four positive outlier hospitals, percent (N=1,222)
Acute myocardial infarction	410	8.1	23.0
Ventricular fibrillation/flutter	427.4	0.7	2.1
Cardiac arrest	427.5	1.6	1.8
Postcardiotomy syndrome/post-op heart failure	429.4	0.9	1.1
Cardiac complications resulting from procedure	997.1	4.2	49.6
Hemorrhage or hematoma complicating procedure	998.1	2.6	9.3
Accidental puncture or laceration	998.2	0.2	2.3
Dehiscence of operation wound	998.3	0.6	1.4
Postoperative infection	998.5	1.4	3.8
Implant of pulsation balloon after date of CABG	37.61	2.1	9.1
Operation for hemorrhage control after date of CABG	39.41, 39.98	0.6	0.3

Source: Derived from discharge abstracts for all patients who underwent coronary bypass grafting using autologous saphenous vein (ICD-9-CM 36.10-36.14, 36.19) in California acute-care non-federal hospitals in 1983 through 1986. Patients under age 18 and those who had valve or congenital heart surgery during the same hospitalization (35.1-35.9) were excluded. A variable was created to indicate the presence of one or more of the related complications specified on this list (plus several low-frequency complications: 996.0, 996.1, 998.2, 998.4, 37.62). A logistic regression model was generated to predict the probability of a complication; these predicted probabilities were aggregated at the hospital level and compared with the observed experience of each hospital (z scores). The nine major hospitals with the most consistently extreme z scores over the 4-year period were assigned to these two groups. Details about the regression model are available from the authors.

if the principal diagnosis is used for a unique purpose such as designating diagnosis-related groups (DRGs), but the first two problems apply to all studies for which administrative data are used. The net effect of these errors in a sample of Medicare discharges from 1984 to 1985 was to increase patients' mean relative weight (a measure of case-mix intensity) by 1.9 percent (Hsia, Krushat, Fagan, and others, 1988). The total overpayment to hospitals amounted to twice the total underpayment. These findings were consistent with earlier regression analyses (Ginsburg and Carter, 1986) which suggested that 2.8 percent of the observed 9.2-percent increase in Medicare case-mix intensity between 1981 and 1984-1985 reflected changes in coding practice induced by the introduction of prospective payment.

Several studies have examined the relative importance of misspecification, miscoding, and missequencing as sources of error. Lloyd and Rissing (1985) attributed 62 percent of the identified errors on discharge abstracts from five VA medical centers to physician error; however, the great majority of these errors (89.2 percent) were missed diagnoses or missed procedures. This type of misspecification is probably less frequent in other current data sets because medical records personnel have assumed the responsibility for listing discharge diagnoses in most non-Federal hospitals. In the National DRG Validation Study (Hsia, Krushat, Fagan, and others, 1988), misspecification of the principal diagnosis was the most common reason for incorrect DRG assignment in a sample of randomly selected Medicare discharges. Hsia (1990) subsequently reported that attending physicians misspecified the principal diagnosis in 53.4 percent of 442 Medicare discharges incorrectly billed as DRG 129 (cardiac arrest).

In outcomes research, these errors may introduce bias through three mechanisms:

1. Incomplete selection of cases: Identification of cases for study depends on accurate coding of the target diagnosis or procedure. If the target diagnosis or procedure is undercoded, the sample may be incomplete and unrepresentative. For example, patients with mild disease or complex associated diagnoses may be at especially high risk of being excluded from a study sample due to undercoding of the target diagnosis.

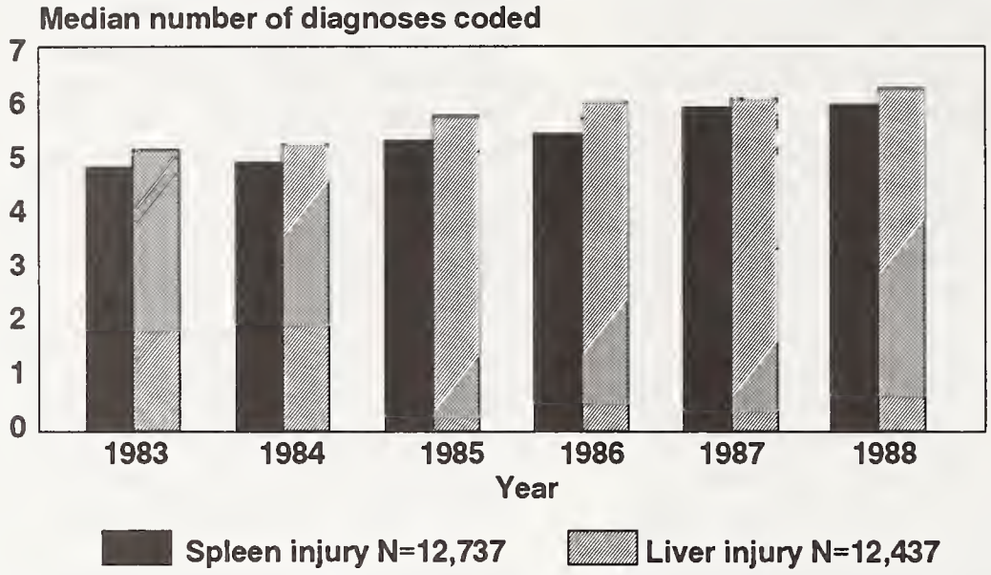
2. Undercoding of comorbidities: Undercoding of comorbidities in studies involving risk adjustment may seriously bias risk estimates. This undercoding may occur when physicians fail to note relevant diagnoses or medical records analysts fail to code all cited diagnoses. Censoring of diagnoses due to the limited number of fields on abstract forms also may cause undercoding because acute problems tend to be listed before chronic comorbidities. This problem is particularly severe in outpatient data sets, which often have room for just one diagnosis per visit. The substitution of acute for chronic diagnoses in patients who experience severe complications can lead to implausible risk relationships (Jencks, Williams, and Kay, 1988). For example, Medicare patients with adult-onset diabetes, ischemic heart disease, or angina listed on their abstracts were significantly less likely to die within 30 days after admission for stroke, pneumonia, MI, or congestive heart failure than patients without these recorded comorbidities. Similar protective effects were reported for several comorbidities in trauma patients in California (Morris, MacKenzie, and Edelstein, 1990).

   Undercoding of comorbidities may bias outcome studies if it varies systematically by hospital type or location. Hospitals that undercode comorbidities may demonstrate poor risk-adjusted outcomes because their patients are actually sicker than the abstracts indicate. Instead of measuring institutional differences in the quality of care, risk-adjusted outcomes may simply measure differences in the quality of coding. Undercoding may also bias time-series studies because the number of diagnoses reported on discharge abstracts has increased steadily over time (Figure 2). This trend incorrectly makes recent patients appear sicker than similar patients in prior years.

3. Overcoding of complications: Medicare's prospective payment system has created a strong incentive to overcode complications that affect medical or surgical management. Many DRGs are stratified according to the presence or absence of complications; the former invariably carries greater weight and thus brings higher reimbursement. Hospitals would therefore be expected to code any complication noted in a patient's chart, no matter how trivial. Simborg (1981) suggested that hospitals are becoming progressively more sophisticated at upcoding DRGs in this manner. Comparisons of hospital outcomes could be biased by systematic variations in the degree of upcoding across institutions.

Approaches to the problem. These problems of incomplete case selection, undercoding of comorbidities, and overcoding of complications all arise from the inadequate reliability and validity of administrative data sets. We can define reliability as the consistency or reproducibility of a measurement (for example, a discharge abstract) across observers and over time. In other words, would two coders reviewing the same chart gen-

Figure 2. Number of ICD-9-CM diagnoses coded on California hospital discharge abstracts, 1983-1988



Source: Derived from analysis of all cases with a principal or secondary diagnosis of liver or spleen injury.

erate the same abstract? Would a coder reviewing a chart this week abstract the same data as he or she would have last week? Validity can be defined as how well a measurement reflects the truth of the phenomenon under study. In other words, does a discharge abstract correctly list all of the relevant diagnoses and procedures? Most approaches to dealing with undercoding and overcoding in administrative data sets involve estimating, comparing, or maximizing their reliability and validity.

Several studies that involved reabstracting randomly selected records have identified factors associated with the reliability of ICD-9-CM coded diagnoses. For example, the Institute of Medicine (IOM) (1977a) reported that diagnosis codes compared at the three-digit level were more reliable than those truncated at four digits. This finding suggests that researchers can minimize bias by using three-digit codes to define comorbidities and complications wherever possible. Unfortunately, the tradeoff between improved reliability and diminished specificity often precludes the use of three-digit codes. For example, chronic and acute renal failure share the same three-digit code even though the former is a comorbidity and the latter may be a complication of therapy.

The estimated reliability of ICD-9-CM coded diagnoses can also be enhanced by ignoring their sequence. In a reabstracting study performed in conjunction with the National Hospital Discharge Survey (NHDS), the principal diagnosis identified by the research team (Demlo and Campbell, 1981) matched that specified on the original NHDS abstract 73.1 percent of the time but was listed as a secondary diagnosis on an additional 7.1 percent of the abstracts. In selecting cases for inclusion in a study, researchers should generally treat all listed diagnoses and procedures equally, rather than relying on the principal diagnosis or procedure. This suggestion is particularly important for data sets that do not explicitly arrange procedures in descending order of importance (for example, MEDPAR). An exception might be made when the condition that is "chiefly responsible for occasioning the admission" is of particular interest.

Reabstracting studies also have shown that some diagnoses are coded with especially poor reliability. These conditions should not be used as inclusion criteria, comorbidities, or complications in outcome studies. The most commonly missed diagnoses in the VA reabstracting study (Lloyd and Rissing, 1985) were the genus of an infecting microbe, diabetes (250.00), alcoholic psychosis (291.8), hemiplegia (342.9), benign hypertension (401.1), angina pectoris (413), coronary atherosclerosis (414.0), chronic obstructive pulmonary disease (496), urinary tract infection (599.0), decubitus ulcer (707.0), and seizure disorder (780.3). The percentage of cases with no discrepancy between the original abstract and the reabstract in the IOM's (1977a) study varied by principal diagnosis, from 30.2 percent for chronic ischemic heart disease (ICDA-8 412.0 or 412.9) to 97.0 percent for hypertrophy of the tonsils and adenoids (ICDA-8 500). The low-reliability outliers in two subsequent IOM studies (Demlo and Campbell, 1981; IOM, 1977b) also included stroke (43.2 percent of cases without discrepancies), congenital anomalies (49.5 percent), and endocrine, nutritional or metabolic disease (26.3 percent). California's reabstracting study

(OSHPD, 1990) enumerated 10 conditions that accounted for 53.5 percent of the overlooked diagnoses in 576 records: chronic respiratory disease, status post (for example, transplant, dialysis, pacemaker), stroke residual, arteriosclerotic heart disease, hypertension, substance abuse or dependence, electrolyte or fluid imbalance, urinary tract infection, cardiac dysrhythmia, and diabetes.

We analyzed the OSHPD reabstracting data in greater detail to estimate the reliability of coding for specific diagnoses. Unlike most administrative data sets, California discharge abstracts include up to 25 diagnoses and up to 25 procedures. The overall level of agreement between the original record and the reabstracted version was 84.7 percent for the five-digit principal diagnosis; that percentage increased to 85.7 percent for a four-digit comparison and 91.2 percent for a three-digit comparison. If the reabstract is viewed as a gold standard, then the sensitivity of coding for a diagnosis is the number of true positives from the original abstract divided by the sum of the true positives and the false negatives. Conversely, the specificity is the number of true negatives from the original abstract divided by the sum of the true negatives and the false positives. Table 3 shows that when the sequence of diagnoses is ignored, most acute conditions related to 1 of the 20 target DRGs were coded with over 90-percent sensitivity. Undercoding is more problematic for chronic or historical diagnoses. Codes with less than 80-percent sensitivity according to these analyses should probably not be used as comorbidities or complications.

Reabstracting studies also suggest that some procedures are commonly not coded; these procedures should not be used as selection criteria for outcome studies. The procedures most frequently omitted from VA discharge abstracts (Lloyd and Rissing, 1985) were spinal taps, endoscopies, cystoscopies, computed tomography of the head, intravenous pyelograms, cardiac ultrasonography, and radionuclide scans. In patients aged 59 and over discharged with a diagnosis of MI in Manitoba, inhalation therapy and multiple blood transfusion were missed more often than they were correctly coded (Roos, Roos, Cageorge, and Nicol, 1982). Demlo and Campbell (1981) reported that abdominal and proctologic surgery were reliably coded on NHDS abstracts, whereas obstetric procedures, diagnostic endoscopy, orthopedic surgery, biopsies, and plastic surgery were not. Our reanalysis of OSHPD's reabstracting data confirmed that blood transfusions, treadmill tests, electrocardiographic monitoring, and inhalation therapy are coded with less than 50-percent sensitivity (Table 4). Overall agreement for the principal procedure (75.4 percent) was less than for the principal diagnosis, largely because some hospitals code no procedures on their nonsurgical patients. A followup survey of all 592 California hospitals (Tagupa, 1991) revealed that over 10 percent of the 357 respondents only "sometimes" code respiratory therapy, physical therapy, arterial blood gas measurement, and electrocardiographic monitoring. Several hospitals commented that the anticipated reimbursement method for a specific procedure influences their coding practices. Clearly, patients who undergo minor procedures cannot be reliably identified from discharge abstracts.

One approach to identifying patients who undergo a procedure coded with questionable reliability is shown in Table 5. We wanted to ascertain all coronary artery bypass patients in California who received a diagnostic cardiac catheterization during the same hospitalization. We created variables representing left-sided cardiac catheterization (37.22 or 37.23), right-sided catheterization (37.21 or 37.23), left-sided angiography (88.53 or 88.54), right-sided angiography (88.52 or 88.54), and coronary angiography (88.55-88.57). The number of cases with different combinations of these procedure sets was carefully reviewed (Table 5). Note that 710 patients had a left-sided catheterization code (with or without right-sided catheterization) but no indication of angiography. Conversely, 258 patients had coronary angiography (with or without left-sided cardiac angiography) but no indication of catheterization. Since these combinations are implausible, we defined diagnostic catheterization cases using all three relevant procedure types: left-sided catheterization or angiography of the left heart or coronary arteries.

A final factor that may affect the reliability of ICD-9-CM coded diagnoses is the type of hospital. Specifically, the IOM (1977a) found that hospitals in urban areas and nonprofit hospitals provided more accurate diagnostic information on their discharge abstracts than hospitals in rural areas and for-profit hospitals. Medical school affiliation and bed size were not consistently associated with reliability. Investigators who are not specifically interested in outcomes at rural hospitals may wish to exclude them. Our examination of California's reabstraction data was limited by the confidentiality of participating hospitals; nevertheless, hospitals vary considerably in how reliably they code patients' principal diagnosis and principal procedure (Figure 3).

The reabstraction studies we have reviewed provide useful warnings about when reliability problems are likely to bias outcome studies using administrative data sets. Investigators working with data sets that have not been as extensively evaluated for reliability as the NHDS and California abstracts can apply the lessons from these studies to their own analyses. Yet reabstrac-

Table 3. Reliability of coding of ICD-9-CM diagnoses among 2,579 reabstracted records from 30 randomly selected hospitals in California in 1988

Diagnosis	ICD-9-CM codes	Sensitivity (percent)	Specificity (percent)
Old myocardial infarct	410	54.5	99.96
Pre-infarction angina	411.1	94.4	99.8
Stable angina pectoris	413.0, 413.1, 413.9	85.4	99.7
Congestive heart failure	428, 398.91, 402.01, 402.11, 402.91	96.2	99.8
Diabetes mellitus	250	93.0	99.8
Hypertension	401-405	90.8	99.8
Chronic obstructive lung disease	491, 492, 496	92.5	98.6
Asthma	493	93.2	99.9
Pneumococcal pneumonia	481	55.6	99.9
Any pneumonia	480-486, 487.0	97.9	99.5
Cancer	141-172, 174-208	90.1	99.96
Acute cerebrovascular disease	433-436, 437.1	97.9	99.8
Old cerebrovascular disease	438	77.3	99.8

Source: Data were derived from the OSHPD Reabstracting Project, in which 2,579 of 2,596 randomly selected medical records from 30 randomly selected acute-care hospitals in California from July through December 1988 were reabstracted by an expert coder. The sampling frame was limited to 20 common medical DRGs. Sensitivity is the percentage of cases with a target diagnosis (according to the reabstracted record) that were identified in the original data set. Specificity is the percentage of cases without a target diagnosis (according to the reabstracted record) that were accurately classified in the original data set.

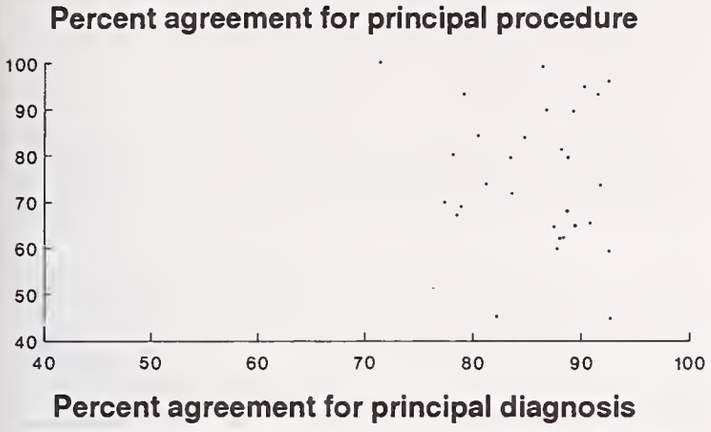
Table 4. Reliability of coding of ICD-9-CM procedures among 2,579 reabstracted records from 30 randomly selected hospitals in California in 1988

Procedure	ICD-9-CM codes	Sensitivity (percent)	Specificity (percent)
Spinal tap	03.31	80.0	100.0
Chemotherapy	99.25	93.9	100.0
Transfusion packed cells	99.04	31.0	100.0
Mechanical ventilation	93.92	92.1	99.96
Respiratory nebulizer therapy	93.94	26.4	99.9
Echocardiography	88.72	66.3	99.9
Electrocardiographic monitor	89.54	39.6	99.7
Cardiovascular treadmill test	89.41	55.6	100.0
Electroencephalogram	89.14	63.9	99.9
Computed tomography, head	87.03	89.8	99.8
Hemodialysis	39.95	90.6	100.0
Flexible colon/sigmoidoscopy	45.23-45.24	91.4	99.9
Barium enema	87.64	66.7	99.96

Source: Data were derived from the OSHPD Reabstracting Project, in which 2,579 of 2,596 randomly selected medical records from 30 randomly selected acute-care hospitals in California from July through December 1988 were reabstracted by an expert coder. The sampling frame was limited to 20 common medical DRGs. Sensitivity is the percentage of cases with a target diagnosis (according to the reabstracted record) that were identified in the original data set. Specificity is the percentage of cases without a target diagnosis (according to the reabstracted record) that were accurately classified in the original data set.

tion studies, as useful as they are, do not really address the issue of validity. It may be reassuring that two coders reviewing the same record in different places generate similar abstracts, but how do we know that either abstract represents what truly happened to the patient? Validity is much more difficult to assess than reliability because there is no clear gold standard. Even the most accurate and precise coder is limited by the accuracy of the medical record. Important comorbidities may not be mentioned, either because a complete assessment of the

Figure 3. Hospital-specific reliability of principal diagnosis and principal procedure, ICD-9-CM coded to four digits



Source: Derived from reanalysis of California hospital discharge reabstracting study. N=2,579 patients at n=30 hospitals.

patient was deferred or because the results were not properly transcribed. Fear of malpractice liability may induce physicians to avoid mentioning certain complications. Physicians or nurses may make incorrect diagnoses, or their provisional diagnoses may be disproved without being overtly repudiated. Notes in the medical record may be illegible or incomprehensible. Reabstraction cannot ascertain the validity of data derived from misleading or incomplete medical records.

Several approaches are available to deal with this problem of unknown validity. The best method is to compare administrative data with independently and prospectively collected data on the same patients. As part of the ischemic heart disease Patient Outcomes Research Team (PORT) study, Jollis (1991) linked data collected by cardiology fellows while they prepared patients for coronary catheterization with discharge abstracts generated by medical records staff. The resulting preliminary sample included 2,125 inpatients with angina due to stable coronary artery disease and no prior history of coronary revascularization who underwent an initial diagnostic catheterization at Duke University Medical Center between 1985 and 1989. Table 6 shows that only a few prognostically significant conditions (for example, diabetes) were consistently coded. The validity of coding for mitral regurgitation (MR) and congestive heart failure was directly related to clinical severity. For example, severe MR was coded with 81-percent sensitivity, whereas mild MR was coded with just 25-percent sensitivity. Roos, Sharp, and Cohen (1991) linked hospital claims for prostatectomy and cholecystectomy from a Winnipeg teaching hospital with data collected preoperatively by attending anesthesiologists. Patients with cardiovascular comorbidity identified from both files appeared to be more chronically ill than patients whose comorbidity appeared only in the anesthesia file.

Although this type of comparison is very useful, most researchers cannot obtain primary data on the cases in

Table 5. Frequencies of various combinations of catheterization-related procedure codes in a sample of 23,199 coronary artery bypass graft patients in California, 1983-1988 (listed in descending order)

Left heart catheterization 37.21, 37.23	Right heart catheterization 37.22, 37.23	Left heart angiocardiography 88.53, 88.54	Right heart angiocardiography 88.52, 88.54	Coronary arteriography 88.55, 88.56, 88.57	Frequency
Y		Y		Y	4,038
Y	Y	Y		Y	3,252
Y				Y	1,379
Y	Y			Y	1,236
Y					366
Y	Y				344
Y		Y			333
	Y				253
Y	Y	Y	Y		212
				Y	210
Y	Y	Y	Y	Y	201
Y	Y	Y			104
	Y	Y		Y	86
	Y			Y	74
Y		Y	Y		67
		Y		Y	48
Y		Y	Y	Y	35

Note: Y indicates the presence of an ICD-9-CM code that represents the specified procedure.

their administrative data base. A simpler alternative is to compare the prevalences of diagnoses in a secondary data set with those reported in a comparable primary source. If the patient populations are similar, then large discrepancies in the prevalence of clearly defined conditions should arouse suspicion. As part of the ischemic heart disease PORT, we have contrasted data from four sites (Romano, Roos, and Luft, 1991), two that obtain discharge abstracts or claims and two that maintain independent registries. Table 7 illustrates that diabetes and several other comorbidities appear to be validly coded in one of our administrative data sets, whereas hyperlipidemias and other asymptomatic conditions do not, when compared with a primary data set that includes similar patients. The prevalences of comorbidities in patients enrolled in randomized trials should not be used as standards for such comparisons. Geographic registries, such as those maintained for cancer and birth defects in many areas, generally provide more representative samples.

Risk ratios derived from secondary data sets also can be compared with those derived from primary or registry data. Undercoding of clinically significant comorbidities may bias estimates of risk in either direction. If the misclassification is random, then the bias is always toward the null; however, differential misclassification may occur if patients with mild forms of a condition have a greater probability of misclassification than those with severe manifestations. In this case, the risk estimates derived from administrative data sets may be biased upward. If the odds ratio or relative risk associated with a specific factor in an administrative data set differs markedly from that suggested by clinical studies, then the factor should probably be ignored or redefined. For example, preliminary data from the ischemic heart disease PORT show that the mortality risk ratio associated with cardiomegaly (ICD-9-CM 429.3) in coronary bypass patients varies from 1.0 in an administrative data set to 4.1 in a primary data set.

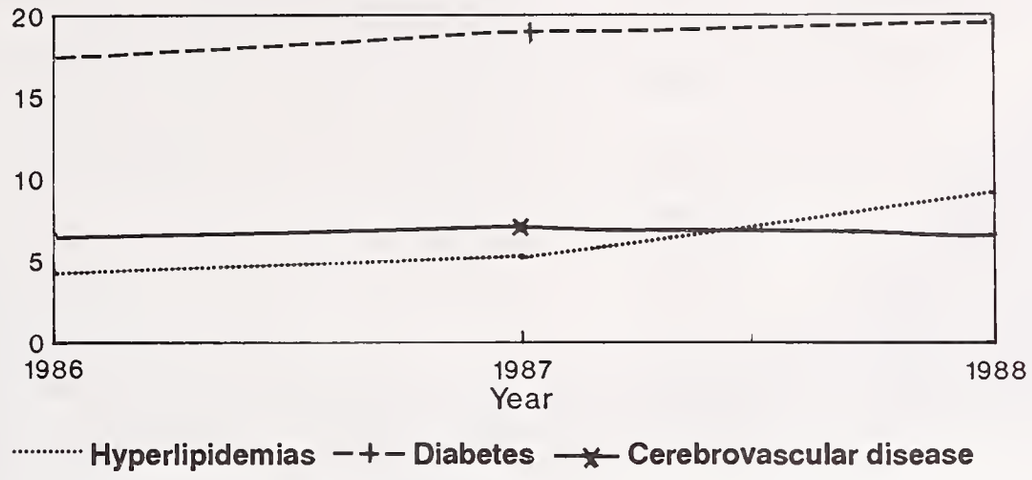
Researchers using data collected over several years should examine temporal trends in the prevalence of important diagnoses. If the prevalence of a comorbidity appears to be increasing dramatically, then undercoding during the earlier period should be suspected. For example, hyperlipidemias (ICD-9-CM 272.0-272.4) were described in only 4.2 percent of CABG patients in California in 1986, but in 9.1 percent of patients in 1988 (Figure 4). This increase contrasts with the stable prevalence of other associated conditions, such as diabetes (17.4 percent in 1986, 19.7 percent in 1988) and cerebrovascular disease (6.4 percent in 1986, 6.6 percent in 1988). Rapid increases in the prevalence of complications more likely reflect overcoding during the DRG era. The list of comorbidities or complications should be modified to exclude those with implausible temporal changes. If such a risk factor is indispensable, then separate predictive models should be developed for different time intervals, or interaction terms between comorbidities and time should be introduced so that the impact of a comorbidity can change over time. Similarly, time can be entered as a covariate in predicting complications that display major temporal changes. This approach can, however, mask true deterioration in the quality of care.

If hospital characteristics are a subject of analysis, then the prevalence of comorbidities or complications in different hospitals should be compared. For example, a comorbidity that is much more prevalent in one hospital than another serving the same community should arouse suspicion. Blumberg (1991) dealt with this problem in the AMI Medicare Mortality Study by eliminating hospitals below the 5th percentile or above the 99th percentile in the hospital-specific frequency of hypertension and subendocardial infarction. The validity of coded comorbidities in these outlier hospitals is debatable. One possible reason for interhospital variation in the prevalence of comorbidities is that some hospitals employ computer programs that enhance or limit the number of listed diagnoses. For example, nearly 50 percent of discharges following coronary bypass surgery in 1983-1986 at one California teaching hospital listed exactly 14 secondary diagnoses.

If prior hospitalizations or outpatient claims can be linked with the index episode of care, then comorbidities can be validated and additional diagnoses can be identified. Roos (Personal communication, March 1992) found that searching inpatient records up to 1 year prior to the date of coronary bypass surgery increased the recorded prevalence of comorbidities in Manitoba patients by 17-230 percent. Matching with outpatient claims might be even more beneficial because some chronically ill patients have no prior hospitalizations. A series of outpatient visits allow more opportunities for a specific chronic diagnosis to be noted on insurance claims. Of course, reasonable caution should be exercised in interpreting diagnoses that appear only once in a longitudinal record. Time-sequenced relationships between diagnoses and procedures can help validate both sets of codes (Roos, Roos, Cageorge, and Nicol, 1982). If physician visits for pregnancy or cervical disease occur after a total hysterectomy, for example, then either the procedure or the subsequent diagnoses were surely miscoded.

Another approach to validating ICD-9-CM coded complications is to use postoperative length of stay or death as a corroborating measure. Most significant

Figure 4. Temporal trends in prevalence of chronic comorbidities among CABG patients in California, 1986-1988



Source: Hospital reports were derived from the annual survey form that all California hospitals are required to complete.

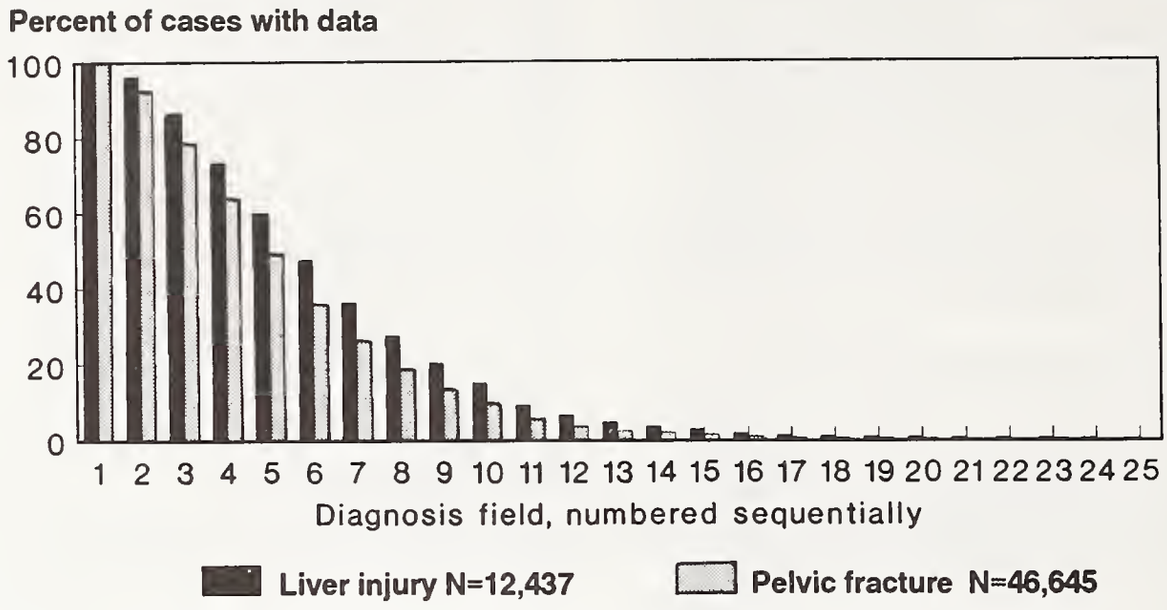
Table 6. Validity of coding of comorbidity diagnoses among 2,125 patients with non-progressive angina due to coronary artery disease who underwent an initial diagnostic catheterization at Duke University Medical Center, preliminary data, 1985-1989

Diagnosis	ICD-9-CM codes	Agreement (percent)	Sensitivity (percent)
Old myocardial infarction	412	75.2	29.1
Angina pectoris	413.0, 413.1, 413.9	25.1	24.6
Congestive heart failure	428, 398.91, 402.01, 402.11, 402.91	89.4	17.1
Cardiomegaly	429.3	84.6	4.9
Mitral insufficiency	424.0, 394.1, 394.2, 396.2, 396.3, 746.6	87.0	29.9
Diabetes mellitus	250	94.4	82.1
Hypertension	401-405	72.9	57.0
Hyperlipidemia	272.0-272.4	71.3	36.7
Peripheral vascular disease or gangrene	440.2, 443.1, 443.81, 443.9, 785.4	89.2	23.2
Cerebrovascular disease	434-436, 437.1, 438	91.0	8.2

Source: Derived from linkage of data collected by cardiology fellows while preparing patients for coronary catheterization with discharge abstracts generated by medical records staff. Agreement is the percentage of cases where both data sources agreed as to the presence or absence of a diagnosis. Sensitivity is the percentage of cases with a target diagnosis (according to the prospective data set) that were also identified in the administrative data set. The number of cases varies slightly for each item. Adapted from Jollis, 1991.

complications prolong patients' postoperative stay. The magnitude of association between an ICD-9-CM coded complication and postoperative length of stay can be estimated through multiple linear regression, controlling for age and other relevant patient characteristics. For example, ICD-9-CM complications such as congestive heart failure and atrial fibrillation did not predict length of stay or hospital charges in AMI patients at a Boston teaching hospital (Barbash, Safran, Ransil, and others, 1987). One can also determine the complicating conditions most strongly associated with death. Sixty percent of patients who died in hospital after a hysterectomy in California between 1983 and 1988 had one or more of the following complications coded: deep venous thrombosis, pulmonary embolus, sepsis, peritonitis, pneumonia, shock lung, disseminated intravascular coagulation, AMI, respiratory failure, cardiac arrest, or acute renal failure. Yet only 31 percent of persons with one or more

Figure 5. Coding of 25 sequential diagnosis fields on California hospital discharge abstracts, 1983-1988


Percent of cases with data
100
80
60
40
20
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
Diagnosis field, numbered sequentially
Liver injury N=12,437
Pelvic fracture N=46,645

Source: Derived from analysis of all cases with principal or secondary diagnosis of liver injury or pelvic fracture.

of these complications died; the rest experienced an adverse but not fatal outcome. Having thereby validated the coding of these complications, one can modify the dependent variable to include either death or severe complications. This change increases the dependent mean threefold and improves statistical power.

As part of the validation process, researchers should estimate the truncation of diagnoses in their administrative data. If 50 percent of patient records have all diagnosis fields filled, then many diagnoses are probably not being coded. Ideally, the number of records with all fields filled should be similar to or less than the number with all but one field filled. For example, California discharge abstracts for major trauma diagnoses demonstrate a smooth "extinction curve" as one scans across the 25 diagnosis fields (Figure 5). Comparable curves derived from MEDPAR data with only five diagnosis fields are clearly truncated.

If these tests of reliability and validity cast doubt on the coding of a particular diagnosis or procedure, then the researcher should proceed cautiously. Questionably coded comorbidities should be ignored or specified in a different manner. Questionably coded complications can be ignored in favor of "harder" outcomes such as death; however, this strategy may be impractical because of poor statistical power if mortality is very low. Furthermore, complications are a clinically important outcome and result in substantial medical costs and long-term disability. It is conceptually just as important to identify factors associated with complications in survivors as it is to identify factors associated with death. The researcher must therefore balance the risks and benefits of alternative approaches to specifying complications.

Inability to Link with Postdischarge Events

Nature of the problem. In some secondary data sets, it is impossible to link serial admissions or visits due to the absence of patient identifiers. This deficiency is unfortunate because rates of rehospitalization or recurrent illness are important outcome measures. Without patient identifiers, one also cannot estimate 30-day or 1-year mortality following an index event. As an outcome measure, inpatient mortality is subject to bias because of interhospital and interstate variation in discharge practices. The availability of skilled nursing beds, the structure of referral networks, financial incentives for early discharge, and community attitudes regarding care of terminally ill patients are just a few of the factors which drive that variation. For example, the discrepancy between inpatient and 30-day mortality for major medical diagnoses is greater in California than in New York due to shorter mean lengths of stay in New York (Jencks, Williams, and Kay, 1988). That discrepancy also varies markedly by procedure; over 90 percent of 30-day post-CABG deaths in California occur in hospital versus just 36 percent of 30-day postprostatectomy deaths (Figure 6).

Approaches to the problem. No truly satisfactory way exists to evaluate postdischarge outcomes without unique identifiers. All approaches to this problem depend on having access to a parallel data set that includes both the outcome of primary interest (for example, 30-day mortality) and the surrogate outcome available

in one's primary data set (for example, inhospital mortality). If hospitals are the primary unit of analysis, the Pearson correlation coefficient between the two outcomes can be calculated across all hospitals in the parallel data set. Similar calculations can be done at the county or zip code level. If the r statistic exceeds some arbitrary value, the surrogate measure is considered valid. This test should be performed for each diagnosis or procedure under study because the correlations among outcome measures may vary (Roos and Brazauskas, 1990). Caution should be exercised in applying surrogate measures to identify hospital outliers, even when the correlation coefficient is close to 1.

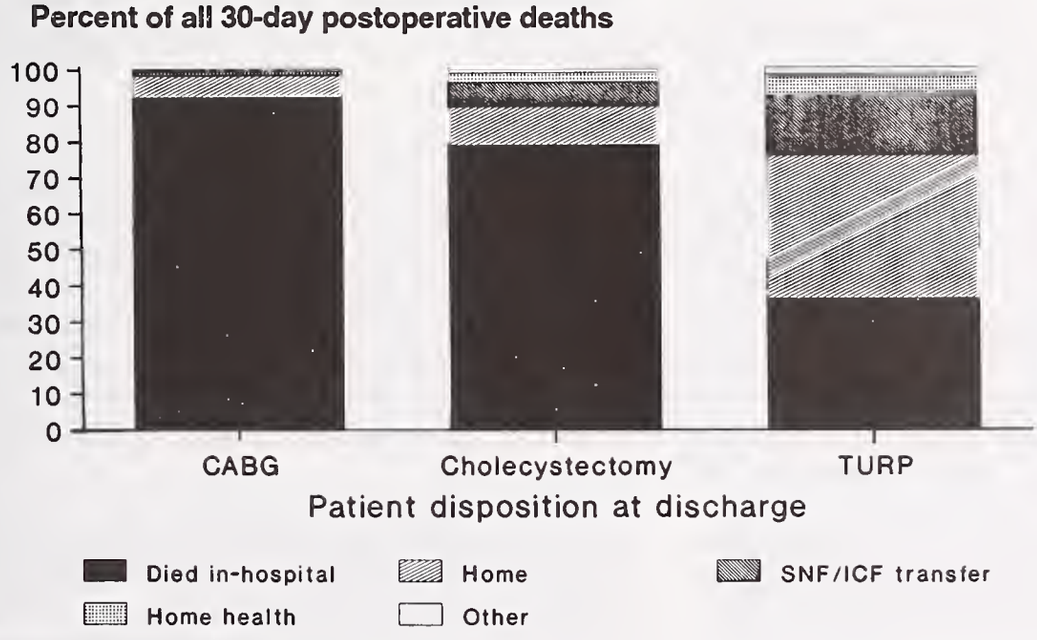
If the correlation coefficient is suboptimal, an adjustment to minimize bias might be considered. Patients discharged to nursing homes or other hospitals often have higher 30-day postoperative mortality that those sent home. Outcome studies should not reward hospitals or geographic units that merely shift their inpatient mortality to other settings. To obviate this problem, high-risk discharges can be coded as some fraction of a death, based on their estimated risk derived from a parallel data set that permits linkage with the National Death Index (for example, MEDPAR). For instance, New York Medicare beneficiaries who underwent cholecystectomy and were discharged to home in 1987-1988 experienced a 30-day mortality of 0.45 percent (74 of 16,354), whereas those discharged to another hospital had a 14.9-percent (3 of 87) risk of death and those discharged to a long-term care facility had a 1.3-percent (9 of 708) risk of death. Through multivariate methods, other patient characteristics that predispose to early postdischarge mortality can be identified in separate data sets; the mortal status at discharge of individuals with these characteristics can be adjusted accordingly.

Timing of Secondary Diagnoses Unknown

Nature of the problem. Many secondary data sets specify the date of each procedure performed during a hospitalization, but few address the timing of diagnoses. This deficiency may lead to confusion between comorbidities and outcomes. For example, AMI may be either an indication for CABG or a complication thereof. If a researcher assumes that an AMI preceded surgery and uses it as a predictor of postoperative mortality, then he or she may explain too much of the variance in outcomes and miss a true quality-of-care effect. On the other hand, a researcher who assumes that an AMI occurred after surgery risks making hospitals that operate on the sickest patients appear to have the highest complication rates.

Approaches to the problem. Differentiating comorbidities from complications is difficult, but several approaches may be useful. If serial hospitalizations or outpatient visits can be linked, then the questionable comorbidity may be used only if it was noted during a prior episode of care. A diagnosis present in a prior record is undeniably a comorbidity. This approach only works for chronic conditions that are unlikely to disappear, such as cerebrovascular or peripheral vascular disease. Acute conditions such as MI may resolve and later recur. The relative rarity of prior admissions within a reasonable time window (for example, 6-12 months) and the

Figure 6. Patient discharge disposition for 30-day postoperative deaths, California 1987-1988



Source: From 1987-1988 pooled QC/MEDPAR data.

Table 7. Frequencies of chronic comorbid conditions among coronary artery bypass graft (CABG) patients from two sites, preliminary data from the Chronic Ischemic Heart Disease PORT group

Diagnosis	ICD-9-CM codes	Site A Prospective (percent)	Site B Administrative (percent)
Old myocardial infarction	410	41.9	14.5
Angina pectoris	411.1, 413.0, 413.1, 413.9	100.0	64.6
Congestive heart failure	428, 398.91, 402.01, 402.11, 402.91	9.3	10.9
Cardiomegaly	429.3	9.4	1.3
Mitral insufficiency	424.0, 394.1, 394.2, 396.2, 396.3, 746.6	18.9	2.9
Diabetes mellitus	250	16.3	18.7
Hypertension	401-405	46.4	31.8
Hyperlipidemia	272.0-272.4	25.2	6.3
Peripheral vascular disease or gangrene	440.2, 443.1, 443.81, 443.9, 785.4	19.3	4.9
Tobacco use	305.1	66.7	1.7
Premature ventricular contractions	427.69	2.7	7.6

Note: Site A represents prospectively collected data for 3,074 patients with stable angina who underwent CABG following a diagnostic catheterization at a major academic medical center between August 1969 and June 1989. Site B represents hospital discharge abstracts for all 55,489 cases of CABG without concurrent open heart surgery in one State between January 1986 and December 1989. Site A excluded patients admitted with acute myocardial infarction or progressive angina, while site B did not.

Source: Adapted from Romano, Roos, and Luft, 1991.

unreliable coding of outpatient diagnoses (Bright, Avorn, and Everitt, 1989) also limit the yield of this approach; however, the CPT-4 coding system encourages precise coding of procedures in outpatient data sets. Outpatient procedures may therefore be useful as indicators of an underlying chronic disease (for example, hemodialysis for chronic renal failure).

A weaker approach is to examine other variables that may clarify the confusion. For example, a patient who suffers an AMI after coronary bypass surgery is likely to have a prolonged postoperative stay, whereas a patient whose MI preceded CABG is more likely to have a preoperative delay.

Other data sets (for example, registry data) can be studied to ascertain whether a specific condition usually precedes or follows surgery. If this analysis reveals that the diagnosis precedes surgery in 90 percent of patients, then it can safely be treated as a comorbidity. Patients in whom the diagnosis followed surgery can be identified and characterized in a primary data set. This characterization may enable one to isolate cases from an administrative data set that follow a pattern consistent with postoperative morbidity. The 5-percent Medicare sample file with linked Part A and Part B records—although inferior to primary data for this purpose—also may permit investigators to determine how often a specific diagnosis precedes the problem of interest in a representative sample of similar patients.

As a last resort, it may be necessary to ignore the offending variable. Surrogate variables can be used instead if their timing relative to the index procedure is clear. For example, it is impossible to identify patients who had a retained common bile duct stone after cholecystectomy based solely on ICD-9-CM diagnoses. Reoperation for exploration of the common bile duct or endoscopic extraction of a stone is a reasonable surrogate for the diagnosis of interest. Procedure variables are preferred because the chronology of the primary and secondary procedures can be clearly delineated.

Conclusion

A wide variety of potential problems are associated with the use of administrative data sets. Although numerous approaches for addressing these problems have been offered (Table 8), none provides a perfect solution. This discussion may therefore raise the question of whether one should pursue research with such inherently problematic data. It is our belief that, in spite of clear limitations, secondary data sets have important roles in studies of appropriateness, quality, and outcomes of care.

The cost of administrative data per case is a small fraction of the cost of prospectively collecting comparable information. More importantly, even in multicenter collaborative studies with large budgets, the participating institutions are few in number and far from representa-

Table 8. Overview of problems and approaches for dealing with large administrative data sets

Problem	Approaches
Sampling frame is of unknown size, incomplete, or unrepresentative of the population of interest	1. Use a denominator that accurately reflects the size of the sampling frame: census-based vs. insurance enrollment-based 2. Exclude from all analyses those patients for whom a given insurer is not the primary payer 3. Search multiple administrative data bases (for example, Part A and Part B claims) to find missing cases 4. Compare estimates of procedure or diagnosis frequency derived from secondary data with tallies maintained or reported by each hospital for separate purposes (for example, licensure, designation as a trauma center)
ICD-9-CM classification scheme is often out-of-date, imprecise, lacking in detail, and fails to distinguish clinically dissimilar diagnoses and procedures	1. Search for associated diagnoses, procedures, or characteristics of the hospital stay that may clarify the meaning of a vague ICD-9-CM code 2. Use procedures as surrogates for the diagnosis of interest if the latter is uninterpretable 3. In defining comorbidities or complications, aggregate specific and non-specific codes that demonstrate similar outcomes 4. Exclude data from hospitals that use non-specific ICD-9-CM codes excessively
ICD-9-CM diagnoses and procedures may be coded with poor reliability, leading to incomplete case selection, undercoding of comorbidities, and overcoding of complications	1. Use three or four digit ICD-9 diagnosis codes and three digit procedure codes, if possible 2. Evaluate both the principal diagnosis and all secondary diagnoses, both the principal procedure and all secondary procedures, to avoid bias due to mis-sequencing 3. Avoid using diagnoses or procedures known to be unreliably coded as inclusion criteria, comorbidities, or complications in outcome studies 4. Treat rural and urban hospitals separately, since their coding practices often differ
ICD-9-CM diagnoses and procedures may be coded with poor validity due to mis-specification by clinicians	1. Compare administrative data with independently and prospectively collected data on the same or carefully matched patients 2. Compare the prevalences of diagnoses in an administrative data set with those reported from a comparable primary source 3. Compare risk ratios for comorbid conditions derived from administrative data with those previously derived from primary data 4. Check for implausible temporal trends in the prevalences of important diagnoses 5. Check for implausible variation among hospitals or geographic units in the prevalences of comorbidities or complications 6. Link sequential hospitalizations or outpatient visits to develop a more complete profile of comorbidities and complications 7. Demonstrate that severe complications are associated with death and long length of stay 8. Estimate the truncation of diagnoses due to the limited number of fields in administrative data
Linkage with post-discharge outcomes, such as rehospitalization or death, may be impossible	1. Estimate the bivariate correlation between the optimal outcome measure and that under study across hospitals in a parallel data set 2. Examine the "true" outcomes of persons with dispositions other than death in a parallel data set that permits linkage with post-discharge events; apply those outcomes to the dispositions obtained from an unlinked data set
Relative timing of secondary diagnoses may be unknown	1. Review serial hospitalizations or outpatient visits to determine the sequence of diagnoses 2. Examine other variables, such as postoperative length of stay or disposition, that may clarify whether a condition was present on admission or developed as a complication of treatment 3. Review other data sets or clinical literature to ascertain whether a specific condition is usually a comorbidity or a complication 4. Use surrogate variables that are precisely dated, such as procedures, instead of the condition of interest

tive. To the extent that we need to understand the variability in medical practice patterns and outcomes across community settings, there is no choice but to use unobtrusively collected administrative data.

Recognizing the problems with administrative data enables investigators to improve the quality of their research. In some instances, data problems can be corrected or avoided. In other instances, correction is cur-

rently impossible, and the cautious researcher may choose to focus on another topic rather than be misled by coding-related biases. Many of the problems we have identified are not inherent to administrative data but reflect existing structures and coding rules. For example, New York now requests that hospitals distinguish diagnoses present at admission from those identified later during a hospital stay. Encryption methods are available that can uniquely transform social security numbers. Reabstracting studies, such as that conducted in California, can be used to improve the training of coders. Over time, therefore, administrative data are likely to improve. Simultaneously, with increased experience, researchers will probably be able to identify and then correct or avoid more of the remaining problems.

References

Agency for Health Care Policy and Research. (1991). *The feasibility of linking research-related data bases to federal and non-federal medical administrative data bases* (AHCPR Pub. 91-0003). Rockville MD: Author.

Barbash, G.I., Safran, C., Ransil, B.J., and others. (1987). Need for better severity indexes of acute myocardial infarction under diagnosis-related groups. *American Journal of Cardiology*, 59, 1052-1056.

Blumberg, M.S. (1991). Potentials and limitations of data base research illustrated by the QMMP AMI Medicare Mortality Study. *Statistics in Medicine*, 10, 637-646.

Blumberg, M.S. and Binns, G.S. (1989). *Risk-adjusted 30-day mortality of fresh acute myocardial infarctions: The technical report.* Chicago: Hospital Research and Educational Trust.

Bright, R.A., Avorn, J., and Everitt, D.E. (1989). Medicaid data as a resource for epidemiologic studies: Strengths and limitations. *Journal of Clinical Epidemiology*, 42, 937-945.

Demlo, L.K. and Campbell, P.M. (1981). Improving hospital discharge data: Lessons from the National Hospital Discharge Survey. *Medical Care*, 19, 1030-1040.

Feinstein, A.R. (1988). ICD, POR, and DRG: Unsolved scientific problems in the nosology of clinical medicine. *Archives of Internal Medicine*, 148, 2269-2274.

Fisher, E.S., Baron, J.A., Malenka, D.J., and others. (1990). Overcoming potential pitfalls in the use of Medicare data for epidemiologic research. *American Journal of Public Health*, 80, 1487-1490.

Fisher, E.S., Baron, J.A., Malenka, D.J., and others. (1991). Hip fracture incidence and mortality in New England. *Epidemiology*, 2, 116-122.

General Accounting Office. (1991). *Medicare: PRO review does not assure quality of care provided by risk HMOs* (HRD-91-48). Washington: Author.

Ginsburg, P.B. and Carter, G.M. (1986). Medicare case-mix index increase. *Health Care Financing Review*, 7(4), 51-65.

Hannan, E.L., Bernard, H.R., O'Donnell, J.F., and Kilburn, H. (1989). A methodology for targeting hospital cases for quality of care record reviews. *American Journal of Public Health*, 79, 430-436.

Hsia, D.C. (1990). Accuracy of Medicare reimbursement for cardiac arrest. *Journal of the American Medical Association*, 264, 59-62.

Hsia, D.C., Krushat, W.M., Fagan, A.B., and others. (1988). Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. *New England Journal of Medicine*, 318, 352-355.

Iezzoni, L.I. (1990). Using administrative diagnostic data to assess the quality of hospital care: Pitfalls and potentials of ICD-9-CM. *International Journal of Technology Assessment in Health Care*, 6, 272-281.

Iezzoni, L.I., Burnside, S., Sickles, L., and others. (1988). Coding of myocardial infarction: Clinical and policy implications. *Annals of Internal Medicine*, 109, 745-751.

Institute of Medicine. (1977a). *Reliability of hospital discharge abstracts.* Washington: National Academy of Sciences.

Institute of Medicine. (1977b). *Reliability of Medicare hospital discharge records.* Washington: National Academy of Sciences.

Jencks, S.F., Williams, D.K., and Kay, T.L. (1988). Assessing hospital-associated deaths from discharge data: The role of length of stay and comorbidities. *Journal of the American Medical Association*, 260, 2240-2246.

Jollis, J. (1991). *ICD-9/Prospective data comparison.* Abstract presented at the meeting of the Ischemic Heart Disease PORT investigators, June 20, 1991, Boston.

Kennedy, G.T., Stern, M.P., and Crawford, M.H. (1984). Miscoding of hospital discharges as acute myocardial infarction: Implications for surveillance programs aimed at elucidating trends in coronary artery disease. *American Journal of Cardiology*, 53, 1000-1002.

Lloyd, S.S. and Rissing, J.P. (1985). Physician and coding errors in patient records. *Journal of the American Medical Association*, 254, 1330-1336.

McMahon, L.F. and Smits, H.L. (1986). Can Medicare prospective payment survive the ICD-9-CM disease classification system? *Annals of Internal Medicine*, 104, 562-566.

Morris, J.A., MacKenzie, E.J., and Edelstein, S.L. (1990). The effect of preexisting conditions on mortality in trauma pa-

tients. *Journal of the American Medical Association*, 263, 1942-1946.

Mullin, R.L. (1985). Diagnosis-related groups and severity: ICD-9-CM, the real problem. *Journal of the American Medical Association*, 254, 1208-1210.

Office of Statewide Health Planning and Development. (1989). *Licensed services and utilization profiles: Annual report of hospitals*, 1988. Sacramento, CA: Author.

Office of Statewide Health Planning and Development. (1990). *Report of results from the OSHPD reabstracting project: An evaluation of the reliability of selected patient discharge data, July through December 1988.* Sacramento, CA: Author.

Romano, P.S., Roos, L.L., and Luft, H.S. (1991). *Assessing the validity of diagnostic information in secondary data sets.* Abstract presented at the meeting of the Ischemic Heart Disease PORT investigators, June 20, 1991, Boston.

Roos, L.L., Roos, N.P., Cageorge, S.M., and Nicol, J.P. (1982). How good are the data? Reliability of one health care data bank. *Medical Care*, 20, 266-276.

Roos, L.L. and Brazauskas, R. (1990). Outcomes and quality assurance: Facilitating the use of administrative data. *Quality Assurance in Health Care*, 2, 77-88.

Roos, L.L., Sharp, S.M., and Cohen, M.M. (1991). Comparing clinical information with claims data: Some similarities and differences. *Journal of Clinical Epidemiology*, 44, 881-888.

Simborg, D.W. (1981). DRG creep: a new hospital-acquired disease. *New England Journal of Medicine*, 304, 1602-1604.

Steinberg, E.P., Whittle, J., and Anderson, G.F. (1990). Impact of claims data research on clinical practice. *International Journal of Technology Assessment in Health Care*, 6, 282-287.

Tagupa, M. (1991). Results of questionnaire. *Discharge Data Review, Office of Statewide Health Planning and Development* (Vol. 9, 3). Sacramento, CA: Office of Statewide Health Planning and Development.

Tunis, S.R., Bass, E.B., and Steinberg, E.P. (1991). The use of angioplasty, bypass surgery, and amputation in the management of peripheral vascular disease. *New England Journal of Medicine*, 325, 556-562.

Winslow, C.M., Solomon, D.H., Chassin, M.R., and others. (1988). The appropriateness of carotid endarterectomy. *New England Journal of Medicine*, 318, 721-727.

Acknowledgments. The authors gratefully acknowledge the collaborative efforts of other investigators from the Ischemic Heart Disease Patient Outcomes Research Team. We especially thank David Pryor, M.D., Lawrence Muhlbaier, Ph.D., Elizabeth DeLong, Ph.D., and James Jollis, M.D., at Duke University, as well as Leslie Roos, Ph.D., at the University of Manitoba. We also appreciate the outstanding cooperation of Eleanor Meux, Ph.D., and Andra Zach, R.R.A., at the California Office of Statewide Health Planning and Development, as well as the programming assistance of Deborah Peltzman. This research was partially supported by the Agency for Health Care Policy and Research (P01H506503) and the Centers for Disease Control (R49/CCR903697).

Using Physician Claims to Supplement Hospital Data

Janet B. Mitchell, Ph.D.

Introduction

Romano and Luft (in this volume) provided us with a comprehensive overview of the potential problems associated with using hospital claims for research purposes. They emphasized the shortcomings associated with diagnostic coding and discussed similar problems with procedure coding. Two primary problems with procedure coding on hospital claims are: (1) not all procedures are routinely coded, and (2) the International Classification of Diseases, 9th Revision (ICD-9) procedure codes themselves are not very specific. Fortunately for the researcher, physician claims can be used to help overcome both of these problems, and that is the subject of this discussion paper.

Incomplete Hospital Procedure Coding

As Romano and Luft noted, a major limitation of hospital claims data is the absence of detailed information on what was actually done to the patient in the hospital. The Part A claims do include information on surgical procedures, but a maximum of only three are retained. The recorded procedures are generally limited to those that affect diagnosis-related group (DRG) assignment, with many diagnostic surgeries missing. Although the medical records abstractor may code nonsurgical procedures as well—such as computerized axial tomography scans—the frequency of such coding tends to be idiosyncratic to individual hospitals.

The only data available on ancillary diagnostic tests are simply charges per revenue center, with no indication of how many or which tests were performed. Also, there is no information on physician visits or consultations or on medical therapies such as kidney dialysis.

Except for some services performed by residents, however, virtually every physician service will show up as a Part B bill. These bills can provide the researcher with an in depth look at the mix of services provided during the hospital stay. An example would be the number and type of cardiac imaging tests administered to acute myocardial infarction patients following admission. Physician bills also can be used to identify the study sample definitively—for example, all patients undergoing cardiac catheterization or all patients receiving bronchoscopy. A further advantage is that physician bills can be used to identify all patients receiving a given procedure, not just those who received the procedure on an inpatient basis.

Unlike hospital abstractors, physicians have an incentive to report every service and procedure they perform; without these itemized bills, they will not get paid. As a result, physician bills should provide a fairly complete picture of the care provided to patients. There is one potential source of missing data, however: services provided by residents in teaching hospitals. The magnitude of this problem is unknown, but it probably is not as great as it may first appear. Although residents are not permitted to bill Part B, teaching physicians are allowed to submit bills even if they did not personally perform the service but only supervised the resident.

Nonspecificity of ICD-9 Codes

Services provided by physicians are billed using the CPT-4 procedure coding terminology, as opposed to the ICD-9 procedure coding system used in hospitals. CPT-4 provides physicians with over 10,000 codes to describe their services. Although this plethora of codes is the despair of many policymakers—who feel it helps fuel physician fee inflation—it is a real boon to researchers!

Unlike the ICD-9 procedure codes that are often vague as to the precise nature of the surgical procedure or diagnostic test, CPT-4 records information in excruciating detail. CPT codes generally provide more clinical specificity along three dimensions: (1) the anatomic location

Dr. Mitchell is President, Center for Health Economics Research, Waltham, MA.

of the procedure, (2) the technical approach used to perform the procedure, and often (3) the clinical indications for the procedure. The ICD-9 coding system includes only a single code (80.2) for all arthroscopies, for example, although a fourth digit may be added to identify the site. By contrast, CPT-4 offers 37 different codes to describe arthroscopy. Although Table 1 displays only 12 of those codes, we can see that they describe not only the joint involved, but the specific procedure (for example, with or without synovial biopsy, with partial versus complete synovectomy), as well as the reasons for the procedure (for example, removal of foreign bodies, infection, debridement of articular cartilage).

A similar example is shown in Table 2. Here, we have a single ICD-9 code for endoscopic excisions in the esophagus, regardless of the precise nature of the endoscopy or the reason for performing it. Using CPT-4 codes, however, we can distinguish endoscopies of the esophagus performed to remove tumors or polypoid lesions from those used to control bleeding or treat esophageal varices.

Table 1. Examples of CPT-4 codes for arthroscopy

29815	Arthroscopy, shoulder, diagnostic, with or without synovial biopsy (separate procedure)
29819	Arthroscopy, shoulder, surgical; with removal of loose body or foreign body
29825	with lysis and resection of adhesions with or without manipulation
29834	Arthroscopy, elbow, surgical; with removal of loose body or foreign body
29835	synovectomy, partial
29836	synovectomy, complete
29837	debridement, limited
29838	debridement, extensive
29871	Arthroscopy, knee, surgical; for infection, lavage and drainage
29872	for infection, lavage and drainage with suction irrigation
29877	Debridement/shaving of articular cartilage (chondroplasty)
29879	Abrasion arthroplasty (includes chondroplasty where necessary) or multiple drilling

Table 2. Comparison of ICD-9 and CPT-4 codes for endoscopic procedures on the esophagus

ICD-9 Codes		CPT Codes	
42.33	Endoscopic excision or destruction of lesion or tissue of esophagus	43235	Upper gastrointestinal endoscopy including esophagus, stomach and either the duodenum and/or jejunum as appropriate; complex diagnostic
By endoscopic approach	Injection of esophageal varices	43243	for injection sclerosis of esophageal and/or gastric varices
	Esophageal polypectomy	43251	for removal of polypoid lesion(s)
	Control of esophageal bleeding	43255	for control of hemorrhage (e.g., electrocoagulation, laser photocoagulation)
	Ablation of esophageal neoplasm	43258	for ablation of tumor or mucosal lesion (e.g., electrocoagulation, laser photocoagulation, hot biopsy/fulguration)

Validity and Reliability of Physician Claims

A question arises concerning the validity of the code assigned by the physician. Romano and Luft (in this volume) did an excellent job of summarizing the research performed to validate ICD-9 codes, but unfortunately very little comparable work has been done to validate what physicians record on their bills. This is probably due in large part to the fact that only relatively recently have researchers begun to work with physician claims. The sheer size of the data bases has tended to discourage their use. For example, a 5-percent sample of Medicare enrollees would generate about 600,000 hospital claims but 34 million physician bills.

A simple rule to evaluate the reliability of data elements on physicians' claims is illustrated in the Figure. Is the data field used for payment? If so, then it is almost always reliable. Examples here would include the billed charge, assignment, and the procedure code itself. Medicare payment amounts, both historically and under the new fee schedule, are determined primarily by the procedure reported by the physician. The physician thus has every incentive to accurately report what he or she did. Inaccurate coding—to report that five grafts were inserted during coronary artery bypass graft surgery, for example, when in fact only two were inserted—is fraud. No one, not even the Inspector General, has ever suggested that such miscoding occurs to any significant extent.

There is one exception, however; anesthesiologists and, to a much lesser extent, assistant surgeons sometimes report different procedure codes from those billed by the primary surgeons. In these cases, my rule has always been to assume that the primary surgeon is right. Since the anesthesiologist may not have actually been in the operating room for the entire procedure, he or she may not remember the precise nature of the surgery. These physicians do not need to have accurate recall, furthermore, because their reimbursement is tied less directly to the specific procedure code. Anesthesia payments are determined not only by the type of surgery but also the operating room time, the presence of any complicating conditions, and delegation of care to certified registered nurse anesthetists (CRNAs).[1]

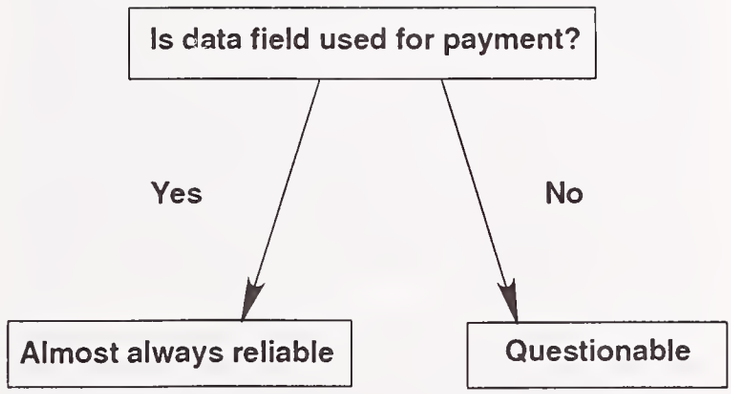
Data fields on physician claims that are not used for payment purposes tend to be less reliable and should be validated whenever possible. One such example is location of service, that is, where the service was performed (office, outpatient department (OPD), or inpatient hospital). In a recent study, we compared costs and outcomes of inguinal hernia repair in inpatient versus outpatient settings. We identified all surgeons' bills for inguinal hernia repair from Part B claims data. We then sought to find matching facility bills for those operations, that is, inpatient Medicare Provider Analysis and Review (MEDPAR) claims, OPD bills, and ambulatory surgical center (ASC) records. As seen in Table 3, we validated the location of surgery as reported by the surgeon for only three-quarters of the hernia repairs. A small number of patients (8 percent) had no matching facility bill and were dropped from the analysis. For the remaining patients, the surgeon and the facility recorded inconsistent locations of service. We chose the facility bills as indicating the correct location; hospitals and OPDs are paid only if the surgery is performed there. Surgeons, however, are reimbursed to the same degree regardless of where the operation is performed. Furthermore, because ambulatory and overnight patients may be mixed together in the same operating room, the surgeon may not be able to discriminate easily between the two.

Managing Physician Claims Data

Because they are so large, most physician claims data bases require careful management; a typical data base might consist of 20-30 million records on two dozen reels of tape (or cartridges). The trick to managing these huge data files, and to keeping computer costs down, is the ability to conceptualize what the final analytic file will look like.

To do this, the researcher needs to decide what the unit of observation or analysis is going to be. Since physician claims are available as line item bills, the researcher must determine if they are to be aggregated and, if so, to what level. Aggregation refers to the process of sum-

Figure 1. Mitchell's maxim for assessing data reliability



[1]Beginning in March 1989, Medicare instituted a requirement that anesthesiologists submit their bills using codes in the anesthesia portion of the CPT-4 manual. Previously, anesthesiologists had used the same procedure codes as surgeons to describe their services. The CPT-4 anesthesia codes are considerably less detailed than the companion surgical codes.

Table 3. Validating location of surgery for hernia repairs

Surgeon's bill	Facility bill	Consistent	Percent of total
Inpatient	Part A inpatient	Yes	55.5
Outpatient	Part A OPD/Part B ASC	Yes	22.6
Outpatient	Part A inpatient	No	6.3
Inpatient	Part A OPD	No	7.4
Inpatient	None	No	8.2

ming up multiple claims based on some shared characteristic, like the patient's identification.

Table 4 provides examples of the different ways claims can be aggregated. Of course, some research questions can be answered by examining the individual line item without any aggregation of claims or complex merging. In this case, the individual service, identified by CPT-4 code, would be the unit of analysis. For outcomes and effectiveness research questions, however, other units of observation are more meaningful. These units might include episodes of illness, individual patients or providers, and geographic areas.

Merging Hospital and Physician Claims

For many outcomes studies, researchers will want to merge hospital and physician claims data. This is easily accomplished with Medicare data by keying on patient identification numbers and the service dates. Since each physician claim includes the date the service was provided, physician bills can be easily assigned to the inpatient stay based on the admission and discharge dates recorded on the hospital claim.

A small percentage of Part A claims may have no associated Part B claims. One source for this nonmatch will be Part A enrollees who are not eligible for Part B coverage (Table 5). Since reliable enrollment data are available from the Health Care Financing Administration, identification of these patients generally can be deduced. Some nonmatches will be health maintenance organization (HMO) enrollees. Certain HMOs, particularly in the Pacific Northwest, are reimbursed on a cost basis for inpatient hospital care but include physician services in their capitated rate. In these instances, a Part A bill is submitted to Medicare, but there are no associated Part B bills.

After hospital and physician claims are merged, bills will be missing for some services that "should" have been provided. In my work, I have found that an attending physician cannot be identified in approximately 5 percent of admissions. (Attending physicians were defined as surgeons for surgical admissions and, in the case of medical admissions, as any physician billing for hospital visits.) "Missing" attendings are somewhat more common in teaching hospitals, suggesting that all services were provided by residents and for patients who died shortly after admission.

Surgical procedures without accompanying anesthesia bills are fairly common, as high as 25-30 percent in some hospitals. This is most frequently found in small hospitals and rural hospitals, where anesthesia is routinely provided by CRNAs working alone. (Until recently, hospital-employed CRNAs were reimbursed under Part A, as a passthrough under the Prospective Payment System. Beginning in 1989, CRNAs were permitted to bill Part B directly for services.)

Table 4. What is the unit of analysis?

Individual service:	Surgical procedure, x-ray, etc.
Episode of illness:	Hospitalization, office encounter, etc.
Beneficiary:	Individual enrollee or patient
Provider:	Physician or hospital
Geographic area:	MSA, county, etc.

Table 5. Sources of non-matches for inpatient episodes

No physician bill:
- Patient not eligible for Part B
- Patient enrolled in HMO

Missing bills:
- No attending physician
  - Services provided by residents only
  - Patient died suddenly
- No anesthesia
  - Anesthesia provided by hospital-employed CRNA

Conclusion

In sum, the careful use of physician claims can overcome many of the limitations of hospital discharge data. These claims can provide more detailed, clinically relevant, and complete descriptors of the services provided to hospital inpatients. At the same time, much more work is needed on assessing the reliability and validity of these claims. This would involve research comparable to that done with hospital records.

Finally, physician claims can also be used to study the outcomes of treatment. Based on the timing of events, some physician services may provide indirect evidence that complications have arisen. An example might be percutaneous transluminal coronary angioplasty, followed closely by coronary artery bypass graft (CABG) surgery. In other cases, the CPT-4 codes themselves can provide direct evidence of complications; an example here would be "exploration for postoperative hemorrhage, thrombosis or infection; chest" (code 35820) following CABG surgery.

Risk Adjustment for Medical Outcome Studies

Lisa I. Iezzoni, M.D., M.Sc.

Introduction

Risk adjustment is an essential methodological ingredient in any study of patient outcomes or the effectiveness of medical care. Its goal is to control for the influence of existing patient characteristics when inferences about the impact of the health care intervention being studied are made. Risk adjustment requires the identification of clinical or other patient attributes (for example, extent of disease or functional impairment) which independently affect "risk" or the likelihood of experiencing specified outcomes, such as death or disability. The underlying assumption of most risk adjustment models is that patients who are more physiologically or functionally impaired will be less able to respond positively to treatment than less compromised patients and are thus at higher risk of a poor outcome regardless of the effectiveness or quality of care.

Few dispute the need to adjust for patient risk in outcome studies: the unresolved question is how to do so. For instance, although investigators attempt to control for patient risk factors in the relatively narrow context of prospective clinical trials evaluating new treatments, even detailed enrollment protocols cannot identify all clinical characteristics that may affect patients' responses to therapy. One goal of the randomized controlled clinical trial (RCT)—the "gold standard" test of relative treatment effectiveness—is to account for a range of patient attributes that could influence therapeutic performance (Meinert, 1986). Randomization aims to assign patients to treatments using a method free of bias, yielding groups that are similar across patient characteristics potentially linked to therapeutic response (both those patient attributes that have been identified and those that are unanticipated, unknown, or unmeasurable). However, because of their high costs and daunting practical impediments, relatively few RCTs have been performed.

Health services research studies of treatment effectiveness differ from prospective clinical trials in several ways that have important implications for the risk adjustment strategy. Most importantly, health services research studies are often retrospective and population based, evaluating events that occurred in the past to groups of patients outside the control of the investigators. Unlike those in neatly packaged clinical trials, patients are not randomly allocated to providers and treatments, and care does not follow prescribed protocols. Identification of patient risk factors thus can become ensnared in attempts to disentangle the appropriateness of services provided and assessments of the technical quality of care. In addition, data are often severely limited, especially information about patients' clinical characteristics. Finally, many health services research studies aspire to evaluate a spectrum of disparate outcomes, ranging from resource use to death to quality of life. The factors that affect risk for these different outcomes may vary widely.

This paper will review several issues raised in risk adjustment for health services research studies of medical effectiveness and outcomes of care, discuss dimensions of patient risk, review other methodological considerations that affect choice of risk adjustment strategy, and describe data considerations. The paper concludes by discussing the limitations of current risk adjustment methods. Examples of specific risk adjustment approaches will be pointed out, primarily those pertaining to medical aspects of risk, such as severity of illness (Table 1).

Dimensions of Patient Risk

"Risk adjustment" has no single, precise definition. It joins a series of other poorly defined words and phrases, such as "case mix," "severity of illness," "sickness," "comorbidity," and "health status," that are increasingly employed not only by researchers, but also by policymakers, financial analysts, assessors of quality of care,

Dr. Iezzoni is an Assistant Professor of Medicine at Harvard Medical School and Research Associate in Medicine, Beth Israel Hospital, the Charles A. Dana Research Institute, and the Harvard-Thorndike Laboratory.

Table 1. Examples of severity measurement methods

Name and acronym	Definition of severity or risk	Data needs
Acuity Index Method (AIM)	Length of hospital stay within DRGs	Discharge abstract
Acute Physiology and Chronic Health Evaluation (APACHE II)[1]	Risk of imminent death, primarily in intensive care	Medical record
Body Systems Count[2]	Count of different organ systems involved	Discharge abstract
Charlson Chronic Disease Index[3]	Risk of death within 1 year of medical hospitalization	Medical record (adapted for discharge abstract)
Computerized Severity Index (CSI)[4]	Treatment difficulty presented to physicians reflected by length of stay	Medical record
Disease Staging Clinical Criteria[5]	Complexity and extent of organ system involvement	Medical record
Disease Staging Q-Stage[6]	Complexity and extent of organ system involvement	Discharge abstract
Disease Staging Q-Scale[7]	Relative resource needs	Discharge abstract
Medicare Mortality Predictor System (MMPS)[8]	Risk of death within 30 days of admission for Medicare beneficiaries in four conditions[9]	Medical record
MedisGroups[10]	Clinical instability	Medical record
Patient Management Categories (PMCs)[11]	Relative cost of services required for effective management	Discharge abstract
Pediatric Risk of Mortality (PRISM) Score[12]	Risk of in-hospital death for patients hospitalized in pediatric intensive care units	Medical record
Rand Measure of "Sickness on Admission"[13]	Risk of death at 30 and 180 days following admission in five conditions[14]	Medical record
Yale Refinement of Diagnosis Related Groups (Refined DRGs)[15]	Relative hospitalization charges, considering complications and comorbidities	Discharge abstract
Uniform Clinical Data Set (UCDS)[16]	Risk for various outcomes potentially related to effectiveness and quality of care	Medical record

[1]Knaus, Draper, Wagner, and Zimmerman, 1985 and 1986; Wagner and Draper, 1984; Wagner, Knaus, and Draper, 1986.
[2]Mendenhall, 1984.
[3]Charlson, Pompei, Ales, and MacKenzie, 1987.
[4]Averill, McGuire, Manning, and others, 1989; Horn and Horn, 1986; Iezzoni, Moskowitz, and Daley, 1989.
[5]Gonnella, 1987; Gonnella, Hornbrook, and Louis, 1984; Gonnella, Louis, and McCord, 1976.
[6]Gonnella, 1987; Conklin, Lieberman, Barnes, and Louis, 1984.
[7]Conklin and Houchens, 1987; SysteMetrics, 1988.
[8]Daley, Jencks, Draper, and others, 1988.
[9]Stroke, pneumonia, acute myocardial infarction, and congestive heart failure.
[10]Brewster, Karlin, Hyde, and others, 1985; Blumberg, 1991; Iezzoni, Ash, Cobb, and Moskowitz, 1988; Iezzoni, Ash, Coffman, and Moskowitz, 1991; Iezzoni and Moskowitz, 1988.
[11]Young, 1984; Young, Swinkola, and Zorn, 1982.
[12]Pollack, Ruttimann, and Getson, 1988.
[13]Keeler, Kahn, Draper, and other, 1990.
[14]Congestive heart failure, acute myocardial infarction, pneumonia, cerebrovascular accident, and hip fracture.
[15]Fetter, Shin, Freeman, and others, 1989.
[16]Vibbert, 1991.

administrators, and managers of health care delivery systems. "Risk adjustment" has no single meaning because of its inextricable linkage to the following question: "risk of what?" This question obviously generates innumerable replies.

Although a uniform definition of what constitutes risk is neither possible nor necessarily desirable, one can suggest sets of patient characteristics that reflect various dimensions or sources of risk. As described in the next section, depending on specific research goals, one or more of these dimensions indicates patient risk and should be considered. Table 2 lists these suggested dimensions of risk, and Table 3 provides brief synopses of 10 clinical scenarios that illustrate the dimensions described below.

A starting point in assessing risk is patient age. In defining other dimensions of risk, some believe that age either has an independent contribution or that the definition of other indicators of risk must be viewed within the context of patient age. The most obvious distinction is among neonates, children, and adults. For example, the MedisGroups severity measure uses different ranges of some physiologic variables (such as hematocrit) to indicate derangements in neonates compared with adults. The Pediatric Risk of Mortality (PRISM) score uses different ranges for such variables as systolic blood pressure and heart and respiratory rates when determining the severity of risk in infants versus children (Pollack, Ruttimann, and Getson, 1988).

The impact of age may depend on the range of ages being studied and the outcome of interest. For example, in an analysis of outcomes of adult patients defined broadly as 18 years of age and older, separating the youngest and oldest persons may be crucial. When persons in narrower age ranges are examined, however, the impact of adjusting for age may be negligible. For instance, when 30-day postadmission mortality for persons 65 years of age and older was examined, age and sex explained only 1 percent of variation in this outcome for congestive heart failure and hip fracture patients, 2 percent for acute myocardial infarction and stroke, and 3 percent for pneumonia (Keeler, Kahn, Draper, and others, 1990). Although the original version of diagnosis-related groups (DRGs)—the basis of Medicare hospital payment—separated some patients into those older and younger than age 70 (Fetter, Shin, Freeman, and others, 1980; Vladeck, 1984), recent revisions of DRGs have not made this distinction, since it was found that separating older and younger patients did not distinguish patients with different hospitalization costs. Most observers, however, believe that the "oldest old," defined variably as above 80 or 85 years of age, are physiologically different from younger patients, even if other factors (such as disease severity) are held constant. One special issue that may need to be considered is the possibility of "ageism" in use of treatments in elderly patients. For example, one study found that, after cancer stage and comorbidities were controlled for, breast cancer patients older than 70 years of age were significantly less likely to receive appropriate therapy than were younger patients (Greenfield, Blanco, Elashoff, and Ganz, 1987).

Acute clinical stability refers to the immediate physiologic functioning of the patient in terms of basic homeostatic measures. This dimension examines the most general and fundamental measures of bodily function (for example, vital signs, serum electrolytes, arterial oxygenation) to evaluate whether the patient will survive or is at imminent risk of dying. Assessments of acute clinical stability are thus most pertinent to studies of acutely ill patients in short time frames. This dimension is independent of diagnosis or "generic," assuming that at this basic level of the "whole person," a "final common pathway" of physiological functions indicates patient risk. For example, Patient F in Table 3 is acutely clinically unstable, with hypotension, a rapid heart rate, diaphoresis, and fever resulting from sepsis and shock. The physiologic compromise of sepsis and shock has immediate implications for short-time survival; at this point, the absence of distant metastases has little bearing on risk of imminent death. A severity measure such as the revised Acute Physiology and Chronic Health Evaluation (APACHE II), based primarily on 12 basic physiologic indicators, reflects this generic perspective on severity of acute clinical stability (Knaus, Draper, Wagner, and Zimmerman, 1985, 1986; Wagner, Knaus, and Draper, 1986).

The principal diagnosis is the leading disease that brings the patient into contact with the health care system and for which services are sought.[1] As noted by Gonnella, Hornbrook, and Louis (1984):

> A disease is a well-defined model of a process of disruption in the normal homeostasis of psychological-physiological systems ... Each disease should be defined in terms of some specific organ or organ system involved ... Every disease must specify some characteristic pathophysiological change ... An etiologic factor or set of factors causing the pathophysiological changes must be given.

[1]The term "principal diagnosis" has a specific meaning within the Medicare program, indicating the condition that, after study, caused the patient's hospital admission. The "principal" is distinguished from the greatest resource use. Because these dimensions of patient risk do not exclusively apply to the inpatient setting, a less specific definition of "principal diagnosis" is intended in this paper, acknowledging that for some patients, it may be difficult to decide which of several serious diagnoses is the "principal."

Table 2. Dimensions of risk

- Age
- Acute clinical stability
- Principal diagnosis ("case mix")
- Severity of principal diagnosis
- Extent and severity of comorbidities
- Physical functional status
- Psychological, cognitive, and psychosocial functioning
- "Nonclinical"attributes (socioeconomic status, race, substance abuse, occupational exposures, culture, religion, ethnicity)
- Health status and quality of life
- Patient attitudes and preferences

Although these criteria yield the most comprehensive specification of disease, it is not always possible to satisfy each requirement, especially determining the etiology; however, even a less rigorous definition of principal diagnosis—such as "low back pain," a condition in which the cause often eludes detection—is frequently considered a critical component of risk adjustment.

Many health services research studies of patient outcomes pertain to a single principal diagnosis, and it is therefore not necessary to adjust for risk based on this dimension (for example, the cases in Table 3 hold the principal diagnosis constant); however, studies that span conditions should generally consider principal diagnosis. In the lexicon of payment policy, adjusting for principal diagnosis is synonymous with adjusting for "case mix." DRGs, a prominent case-mix measure, use principal diagnosis to categorize cases into groups that not only generate similar costs but also are supposedly clinically homogeneous. Principal diagnosis is also important in comparing risks of more clinical outcomes (Green, Passman, and Wintfeld, 1991). For example, in general, risk of imminent death is higher for patients with acute myocardial infarction than for those with acute bronchitis. When the outcomes of a procedure that may be performed for a variety of diseases are evaluated, adjusting for principal diagnosis may be crucial. For instance, when the long-term caseload mortality of two providers performing lobectomies is compared, it is important to separate those procedures performed for lung cancer from those done for other reasons (such as bronchiectasis or vascular malformation).

Table 3. Ten clinical synopses: Adenocarcinoma of the colon

Patient A	No gastrointestinal symptoms. Occult blood found in stool, leading to detection of a small polyp containing adenocarcinoma. Considered a surgically curable lesion. Otherwise in good health.
Patient B	No gastrointestinal symptoms. Occult blood found in stool, leading to detection of a small polyp containing adenocarcinoma. Considered a surgically curable lesion. History of stroke 1 year ago. Paralyzed on right side of body and cannot speak.
Patient C	No gastrointestinal symptoms. Occult blood found in stool, leading to detection of a small polyp containing adenocarcinoma. Considered a surgically curable lesion. Patient homeless with a history of chronic schizophrenia. No known family or friends.
Patient D	Large colonic adenocarcinoma with deep invasion into bowel wall. No evidence of distant metastases. Otherwise in good health.
Patient E	Large colonic adenocarcinoma with deep invasion into bowel wall. No evidence of distant metastases. History of poorly-controlled essential hypertension.
Patient F	Large colonic adenocarcinoma with deep invasion into bowel wall, causing bowel obstruction and perforation. Patient acutely septic and in shock.
Patient G	Widespread metastatic disease. Patient desires active intervention, including intensive care unit admission and intubation if necessary.
Patient H	Widespread metastatic disease. Patient requests "do not resuscitate" status, desiring "comfort measures only."
Patient I	Patient doing well during admission, but dies unexpectedly.
Patient J	Patient doing poorly during admission, and dies as expected.

Some believe that other dimensions of patient risk should be viewed within the context of the principal diagnosis. For instance, in the Computerized Severity Index (CSI), the principal diagnosis dictates the importance assigned to other patient characteristics (Iezzoni, Moskowitz, and Daley, 1989). The same data element may contribute differently to severity depending on the principal diagnoses (for example, a temperature of 101° F is given a higher severity in leukemia than in bronchitis). In addition, the CSI sometimes varies the contribution of unrelated secondary diagnoses to overall severity depending on the principal diagnosis. For instance, secondary diagnoses have a greater impact on overall patient severity ratings for colon cancer than for lung cancer.

Severity of the principal diagnosis is another dimension associated with patient risk, but little consensus exists about its specific definition. According to Gertman and Lowenstein (1984), "severity is what sociologists term a 'folk wisdom' word like 'satisfaction' or 'happiness,' operationally indefinable in a way that is perfectly acceptable to all parties." As described in the next section, severity must be defined within the context of the research questions of a particular study and then operationalized within the limitations of the available data and the window of observation. The importance of this dimension, however, is obvious on an intuitive level (Dubois, Rogers, Moxley, and others, 1987; Green, Wintfeld, Sharkey, and Passman, 1990; Horn, Bulkley, Sharkey, and others, 1985; Iezzoni, 1989). For example, patients A and D in Table 3 differ in the severity or extent of their principal diagnosis, colon cancer. Patient A's lesion is potentially curable, whereas patient D is likely to do poorly with the illness which has deeply invaded the bowel wall. Any study of treatment effectiveness for colon cancer must distinguish patient types A from D.

The extent and severity of comorbidities is an additional dimension that must be contemplated. Comorbidities are diseases unrelated in etiology to the principal diagnosis. For example, cerebrovascular disease is an important comorbidity for patient B. The impact of a comorbidity may depend on the time frame of observation. The stroke of patient B likely will have a more immediate impact on patient longevity than the hypertension of patient E. Although many agree that comorbidities are crucial for predicting long-term survival (Charlson, Pompei, Ales, and MacKenzie, 1987), they also appear to be important when such shorter term outcomes as hospital mortality are examined. As Greenfield, Aronow, Elashoff, and Wantanabe (1988) reported in a California study of hospital mortality, the three hospitals in their study with the highest comorbidity levels were also considered high-mortality outliers.

Comorbidities can affect patients' ability to tolerate treatments and respond to therapy. In particular, comorbidities are often the crucial determinants of risk for operative complications (Goldman, Brown, Levy, and others, 1982; Jewell and Persson, 1985; Savino and Del Guercio, 1985; Schneider, 1985). For example, one clinical index of operative risk predicts the likelihood of cardiac complications for patients having noncardiac surgery, assigning points for generally comorbid conditions, such as myocardial infarction within 6 months, signs of chronic liver disease on physical examination, and laboratory tests indicating chronic pulmonary disease and renal failure (Detsky, Abrams, McLaughlin, and others, 1986; Goldman, 1983; Goldman, Caldera, Nussbaum, and others, 1977). Patients with significant disease involving multiple organ systems are also less likely to tolerate medical treatments, such as chemotherapy, and to exhibit the therapeutic response of patients without a heavy burden of coexisting illnesses.

For the sake of clarity, this discussion has separated severity of the principal diagnosis and comorbidities into different dimensions of risk. When forced to quantify these dimensions, however, many observers merge them into an overall assessment of patient severity of illness or "sickness." For instance, the American Society of Anesthesiologists Physical Status Classification preoperative evaluation indicates the risks of perioperative death on a five-point scale encompassing all aspects of the patient's presentation. Despite the assignment of specific severity scores to each individual diagnosis, the CSI algorithm also computes an overall severity score for the patient, combining the impact of all diseases present. MedisGroups considers all "key clinical findings" regardless of the patients' diagnoses or etiologies of the abnormalities in creating its global admission severity score (Brewster, Karlin, Hyde, and others, 1985; Iezzoni and Moskowitz, 1988).

Another important dimension of patient risk is physical functional status. As with acute clinical stability, functional status represents another "final common pathway" of diseases: for example, the inability to walk may result from innumerable factors. Functional status is known to be associated with long-term patient outcomes and sense of well-being (Stewart, Greenfield, Hays, and others, 1989); however, functional status indicators could also be important predictors of outcomes in shorter terms, such as individual hospitalizations. In developing the Medicare Mortality Predictor System (MMPS), Daley, Jencks, Draper, and others (1988) found that ambulation (scored as able to walk, requiring assistance to walk, and unable to walk) was a significant predictor of mortality 30 days after admission for pneumonia or congestive heart failure. According to Keeler,

Kahn, Draper, and others (1990), the inability to walk before admission was highly predictive of 30-day mortality for pneumonia patients.

A related dimension is psychological, cognitive, and psychosocial functioning. This dimension is particularly important when outcomes are examined outside of controlled institutional environments such as the hospital, where staff purportedly minister to all patient needs (for instance, give medications in proper dosages and on schedule). In the community, psychological and cognitive factors can compromise patients' activity levels, self-care, motivation, and perceptions in ways that could have serious consequences for their response to treatment. For example, a patient suffering severe cognitive deficits from Alzheimer's disease will be unable to care for himself/herself independently. Marital status and social supports also may be important: one study suggested that unmarried men may have higher cancer death rates than their married counterparts (Jenkins, 1983).

A variety of "nonclinical" patient attributes, often placed outside the purview of traditional clinical measures, may nonetheless be critically important as predictors of poor outcome (Breslow, 1989). Low socioeconomic status has been linked to excess death rates (Otten, Teutsch, Williamson, and Marks, 1990), perhaps due to delayed access to medical services or other fundamental chronic deprivations that subtly impair physiological health. For example, cancer survival rates for poor people are 10-15 percent lower than for others (American Cancer Society, 1989). Patients living in hostile settings (for example, the homeless, patients in abusive or neglectful home environments) may fare less well. Race can be important: for instance, even after controlling for many characteristics, including stage of disease, researchers found higher fatality rates for certain cancers in black Americans than in other persons (Manton, Patrick, and Johnson, 1987). Such compromising social factors as alcohol and substance abuse also have an impact on patient outcomes. Occupational exposures may actually cause or increase the susceptibility to certain illnesses. In addition, culture, religion, and ethnicity may be related to outcome by affecting compliance with therapy, diet and other attributes of daily life, and attitude toward health and medical care. A well-known example is the prohibition of Jehovah's Witnesses against having blood transfusions, even when potentially life-saving.

A number of measures have tried to meld these many dimensions into an overall assessment of health status and quality of life (Lohr, 1989; Spilker, Molinek, Johnston, and others, 1990). Although a global assessment of health status is often used in health services research as the outcome of interest, it can also be an important indicator of risk (Brook, Ware, Davies-Avery, and others, 1979; Steinwachs, 1989). Health status measures generally rely on extensive data obtained directly from patients, encompassing a range of dimensions such as severity of illness, physical capabilities, psychosocial functioning, sense of well-being, and "quality of life." Despite their intrinsic appeal, health status measures have not been used widely in health services research or effectiveness measurement, in part because of the costliness of data collection and questions involving their conceptual foundations (Deyo and Patrick, 1989). New patient outcome studies, however, are increasingly viewing health status measures as more sensitive and meaningful indicators of outcome than, for example, death.

Finally, patient attitudes and preferences for outcomes are very important when these outcomes are contemplated. Some patients are more likely to seek aggressive care than others (for example, patient G in Table 3). Such patients will be more likely to suffer iatrogenic complications than those refusing aggressive interventions (for example, patient H), but they may delay death or disability. Patients who participate more actively in decisionmaking about their care have been shown to do better (Kaplan and Ware, 1989). A special case reflecting patients' goals and wishes is the use of "do not resuscitate" (DNR) orders. Not surprisingly, Daley, Jencks, Draper, and others (1988) found that admission orders of DNR were highly predictive of death for stroke, pneumonia, and acute myocardial infarction in addition to the severity indicators, and they included it in the MMPS for these three conditions. Zimmerman, Knaus, Sharpe, and others (1986), however, investigated DNR orders in intensive care units at 13 hospitals, and they found statistically significant variations in rates of DNR use unexplained by patient age, prior heath status, diagnosis, or severity of illness. Wachter, Luce, Hearst, and Lo (1989) examined use of DNR orders at three teaching hospitals and found that patients with acquired immunodeficiency syndrome (AIDS) or lung cancer were significantly more likely to receive DNR orders than patients with cirrhosis or severe congestive heart failure, despite similar prognoses.

Methodological Considerations Influencing Risk Adjustment

Four closely linked methodological decisions have particular implications for the risk adjustment strategy:

1. Unit of analysis;
2. Time frame;
3. Outcome of interest; and

4. Nature of the treatments or interventions being evaluated.

The unit of analysis or observation can be defined in different ways, such as a particular disease, an individual patient, a group of patients (for example, those cared for by a particular hospital or provider), a provider, or a group of providers (Whiting-O'Keefe, Henke, and Simborg, 1984). Different risk adjustment strategies have been designed for different units of observation. For instance, clinical disease staging assigns a severity score from 1 to 4 to each disease experienced by a patient; thus, a patient could have multiple severity ratings (Gonnella, Hornbrook, and Louis, 1984). This version of disease staging does not compute an overall severity score for the patient. A measure that aims to adjust for individual patient risk must be powerfully predictive at this individual level. For example, some observers propose making triage decisions to limited intensive care unit beds based on a patient's APACHE II score (an indicator of level of risk of imminent death). Most of the measures shown in Table 1 were intended to be used across groups of patients and are not well suited to making inferences about individual patients.

The choice of timeframe, or length of observation, is closely related to the definition of the unit of analysis. Various options include the hospital admission, a fixed period initiated by a hospitalization (such as 30 or 180 days after admission) or by diagnosis of the condition, an episode of care (such as a time interval or series of services surrounding an index occurrence or intervention), or an interval of years. The time period has important implications for which dimensions of risk must be weighed. For example, if the time period is hours or days, patient F in Table 3 might be more likely to die than patient H, but if patient F is stabilized and the perforation repaired, his or her longer term prognosis is probably better than for patient H. Similarly, in the short term, patients A, B, and C may have comparable risks of death, but in the longer term, patient A is clearly at less risk than B and C.

In general, acute clinical stability and acute attributes of the principal diagnosis and coexisting conditions are critically important components of risk for studies involving short timeframes. In contrast, chronic disability, physical functioning, and various nonclinical factors increase in importance as the time window expands. Most of the currently available severity measures (see Table 1) focus on the short term—the hospital stay or 30 days after admission. One exception is the chronic disease measure developed by Charlson, Pompei, Ales, and MacKenzie (1987), which aims to predict death within 1 year of an index hospitalization for patients admitted to a medical service. The prototype of longer term risk assessments is the clinical staging rubric of oncology: criteria for cancer staging are often delineated based on risk of 3- to 5-year survival. Relatively little systematic information is available about sources of long-term patient risks outside this narrow context of cancer longevity.

The nature of the outcome of interest has significant implications for the risk adjustment strategy. An important distinction is between resource use and more clinical outcome measures. In particular, "severity" may have different meanings when financial costs versus risk of imminent death are evaluated. For example, patient A (see Table 3) may receive an extensive evaluation, including radiological studies, colonoscopy, and various other tests to determine whether the cancer has spread; the patient is then hospitalized for major bowel surgery. This diagnostic investigation and treatment will be expensive, but after all, the patient is "cured." In contrast, patient H has requested "comfort measures only," which may involve relatively inexpensive home or hospice care and a morphine drip for pain control. This patient, whose care is less costly, is clearly at high risk of imminent death.

Different measures have been developed for different purposes (Table 1). Some measures aim to predict resource use (generally the length of stay or cost of a hospitalization), whereas others focus on risk of death. It is not surprising that these measures do not perform well at tasks for which they were not designed. For example, the DRGs are not good predictors of in-hospital death (Iezzoni, Moskowitz, and Ash, 1988), and the MedisGroups admission severity score is not a good predictor of hospitalization costs (Iezzoni, Ash, Cobb, and Moskowitz, 1988; Iezzoni, Ash, Coffman, and Moskowitz, 1991b). Preliminary work suggests that the clinical predictors of hospitalization costs and mortality are different (Iezzoni, Moskowitz, and Ash, 1988), as are the resultant variable weights and models. These conceptual and empirical observations suggest that different risk adjustment strategies may be necessary for different outcomes.

None of the measures in Table 3 aim to assess risk of other types of outcomes, such as complications of care, functional impairment, quality of life, or global health status. For the latter three outcomes, many believe that the optimal predictive model is ultimately derived from a tautology: for example, the best predictor of subsequent functional status is prior functional status. In these instances, information from the risk assessment may appear on both sides of the predictive equation, distinguished only by the time of observation. Obviously,

this analytic structure presents some methodological challenges.

Finally, the nature of the intervention being studied has some implications for the risk assessment strategy. The treatment itself may present its own risks. For example, major surgery generally offers more immediate risks than many medical treatments. In many circumstances, surgery will not be an option for patients with extensive disease in major organ systems, such as the lungs and heart. Thus, it is important to consider serious coexisting diseases when surgical and medical therapies are compared. Similarly, chemotherapy for cancer is a medical treatment that often raises its own innate risks that may be particularly problematic given certain comorbidity scenarios (such as patients with cardiomyopathy or renal failure). In contrast, treatments such as speech or occupational therapy present few intrinsic clinical risks, and therefore, the risk assessment strategy may focus primarily on the indication for the service.

Implications of Data Requirements

From a practical perspective, the most important attribute of a risk adjustment strategy involves its data requirements. The three major data sources in health services research are:

1. Computerized discharge abstract and data files (administrative data);

2. Data elements generally abstracted retrospectively from the medical record or some other source obtained from the health care provider (medical record information); and

3. Primary data collected directly from patients either through interviews or questionnaires (patient-derived data).

Strategies that rely exclusively on administrative data are indisputably less expensive and can encompass thousands if not millions of patients. Despite this, they may not offer the clinical credibility of strategies based on medical record information or patient-derived data. Because of high costs and significant "human subject" and logistical factors, most health services studies focus on the first two types of information, which are reviewed below. An increasing number of patient outcome studies, however, are attempting to collect information directly from patients.

Administrative data. Administrative data are generally produced by health care providers for billing or other payment-related purposes, not medical outcomes studies. Retaining a sense of this financially motivated origin is essential to careful use of these data. The most common forms of such data are the "discharge abstract" (the Uniform Hospital Discharge Data Set) and billing records, such as the Uniform Billing Form (UB-82) required by Medicare (U.S. Congress, Office of Technology Assessment, 1985). The clinical information contained in these hospitalization records generally includes such variables as patient age, sex, discharge status (alive/dead), admission type (such as urgent versus elective), and a small number of diagnosis and procedure codes. Records for outpatient services, which may include primarily procedural information, may be very limited.

For risk adjustment, key features of the administrative files are diagnosis and procedural data. The standard diagnostic lexicon is the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM). ICD-9-CM is the most recent in a series of disease nomenclatures spanning over a century; the next version, the 10th revision, is scheduled for initial dissemination throughout the United States in 1995 (Prospective Payment Assessment Commission, 1990). The current volume contains over 10,300 codes, including codes for "problems" that are not diagnoses, many more symptoms, and "other" reasons for seeking health care (Slee, 1978). The clinically modified ICD-9-CM contains many descriptive terms, potentially enhancing the ability of the nomenclature to describe patients; however, a fundamental question about ICD-9-CM concerns the clinical meaning of its codes. For example, ICD-9-CM includes over 30 codes for different types of anemia, but it does not specify what hematocrit level permits this diagnosis. Because of the absence of clinical definitions, the assignment of ICD-9-CM codes is highly variable, even resulting in regional variations in coding (Feinstein, 1988).

Despite its large number of codes, ICD-9-CM does not include numerous indicators of patient risk (McMahon and Smits, 1986). It does contain codes that suggest clinical instability (for example, hypotension, shock, respiratory failure), codes that reflect disease severity (for example, thyrotoxicosis with thyroid storm, status asthmaticus, hepatic coma), and many codes that could represent comorbid disease; but ICD-9-CM offers little about such risk dimensions as physical or psychological functioning and socioeconomic factors. Both the lack of definition and absence of important prognostic information hinder creation of complete clinical pictures of patients based on their ICD-9-CM codes. Because few administrative files indicate which conditions were present on admission, it is often impossible to differentiate preexisting diseases from complications that may have arisen during the hospitalization.

Closely linked to the nomenclature itself are its coding guidelines (American Hospital Association, 1979).

ICD-9-CM codes are usually assigned to discharge abstracts based on the language used by the physician in listing discharge diagnoses; for Medicare beneficiaries, current Federal regulations require that physicians sign a statement attesting to the accuracy of the diagnoses listed. Despite this, questions persist about ICD-9-CM coding accuracy. Studies performed by the Institute of Medicine (1977) found a high incidence of problematic diagnostic coding. More recent studies have also noted high rates of questionable coding (Hsia, Krushat, Fagan, and others, 1988; Lloyd and Rissing, 1985). The financial motivation for diagnostic coding arising from DRG-based payment is one source of concern. Simborg (1981) cautioned about "DRG creep: a new hospital-acquired disease"—the assigning of more complex diagnoses than warranted or juggling the sequence of codes to maximize reimbursement. The case-mix index (the average DRG payment weight of a hospital's discharges) has increased almost 26 percent, rising from 1.048 in 1981 to 1.319 in fiscal year 1989 (Prospective Payment Assessment Commission, 1991). Although many factors contributed to this increase, changes in diagnostic coding played a role. A study by the Office of the Inspector General, Department of Health and Human Services, found that 61.7 percent of coding errors financially favored the hospital (Hsia, Krushat, Fagan, and others, 1988).

Another difficulty for risk adjustment concerns the extent of coding of secondary diagnoses. Administrative files limit the number of diagnosis codes that can be recorded. For example, Medicare permits coding of up to five diagnoses. Although five diagnoses may be more than adequate to describe an uncomplicated admission, they may not be sufficient for patients with multiple comorbidities. This situation is compounded by the numerous diagnostic codes available to describe diseases and their various manifestations. Almost half of 1987 Medicare discharges had all five diagnoses listed (Steinwald and Dummit, 1989). In a study using Medicare data, Jencks, Williams, and Kay (1988) found that fewer patients with secondary diagnoses such as diabetes mellitus, unspecified anemia, essential hypertension, old myocardial infarction, and angina died than those without these diagnoses. They suggested this counterintuitive finding resulted from coding bias in the listing of secondary ICD-9-CM codes:

> Medicare's limit of reporting five diagnoses would be more likely to truncate chronic diagnoses from the diagnosis list ... Such a bias is plausible because there are numerous acute diagnoses ... that are likely to take precedence when a patient dies or is very ill at the time of discharge (Jencks, Williams, and Kay, 1988).

Procedure coding is complicated because different nomenclatures are used for different purposes. For example, hospitals report procedures using the ICD-9-CM for Medicare, while physicians use the Health Care Financing Administration (HCFA) Common Procedure Coding System (HCPCS). The codes in these two systems do not readily link. For example, the HCPCS provides much more detail about technical aspects of a procedure, and the ICD-9-CM permits more insight about potential quality problems (for instance, ICD-9-CM contains a code for "suture of laceration of uterus" and HCPCS does not). Occasionally, insurance carriers will create their own modification of a basic lexicon, such as the *Current Procedural Terminology* (American Medical Association, 1992), and it becomes even more difficult to obtain consistent information.

Finally, although their availability and size make administrative data immediately appealing, a number of additional structural attributes have implications for their use in risk adjustment. The principal question is whether it is possible to link information from service use at the individual patient level, especially services across settings of care (for example, hospitals, private doctors' offices, nursing homes, homes). Such linkage presupposes that a unique patient identification number is available regardless of payer, a situation which is true for Medicare data bases but few others. With such linked data, it would be possible to identify conditions present at the time of prior service use and thus considered to be chronic. One can also look at patterns of diagnoses over previous hospitalizations or health care encounters to draw inferences about the course of illness. In data bases without such unique patient identifiers, analyses are limited to individual health care encounters, such as hospitalizations, as the units of observation. Beneficial historical information is not available to support risk adjustment.

Issues raised by use of medical record data. Recently, medical records have expanded to the extent that they sometimes include multiple volumes and hundreds of pages concerning a single admission. Despite this, questions remain about the reliability, objectivity, and completeness of the clinical information, especially in the nonhospital setting. One problem is that the putative "objective" findings from diagnostic testing and physical examinations often depend on subjective clinical judgment and are thus susceptible to large interobserver variability. Such poor interobserver reliability commonly occurs even for many tests that appear technologically sophisticated: "Uncertainty creeps into medical practice through every pore ... and the ambiguities grow worse as medical technology expands" (Eddy, 1984). Numerous reports exist of physicians disagree-

ing among themselves about the presence of a physical finding or the interpretation of common tests, such as chest radiographs and electrocardiograms (Anderson, Hill, and Key, 1989; Gjorup, Bugge, and Jensen, 1984; Johnson and Carpenter, 1986; Raftery and Holland, 1967; Wiener and Nathanson, 1976). For example, one study of microscopic cytological examinations from women with cervical abnormalities found that pathologists agreed with each other only 51 percent of the time and with themselves only 68 percent of the time (Eddy, 1984).

Other questions involve the completeness, accuracy, and validity of medical record information, although relatively few studies have been published on this topic (Romm and Putnam, 1981; Starfield, Steinwachs, Morris, and others, 1979). What is reported suggests that the medical record is sufficient for certain purposes (Kosecoff, Fink, Brook, and Chassin, 1987) but often is an incomplete or inaccurate representation of the care provided or of certain aspects of patients' diseases. For instance, Feigl, Glaefke, Ford, and others (1988) reviewed both inpatient records and outpatient radiation therapy charts for patients with certain cancers. Consultations, laboratory and diagnostic test results, and specific inpatient treatments were well-documented in the inpatient record, but important outpatient information (such as mammography results) was poorly documented.

One important factor is whether the facility is involved in teaching. Trainees tend to leave copious notes, although their accuracy is not always assured. The information needed for risk adjustment thus may be more available at training institutions, leading to the possibility of bias in the risk judgment by facility teaching status. For example, patients at tertiary teaching hospitals, with around-the-clock teams of physicians ranging from interns to specialist consultants, could generate multiple physician notes, while an identical patient at a small community hospital could receive only one daily physician note. This documentation difference may be particularly important for clinical findings that tend to wax and wane, such as wheezing or S3 gallop heart sounds.

One problem exacerbated in the last decade is the use of the medical record as a vehicle for cost-containment and, increasingly, as a legal and administrative document. As Burnum (1989) noted, "The loss of confidentiality of medical records has caused physicians and patients alike to withhold clinically important, but sensitive, personal information from medical records that might harm the patient if brought to public view." Affected areas include documentation of psychological disorders, cognitive dysfunctions, substance abuse, sexual practices, and AIDS. To justify a patient's continued hospitalization, "chart notes have become increasingly gloomy." Language is crafted to meet expectations of the varied parties who now gain access to the medical record (for example, insurers, billing offices, lawyers, patients).

Another issue involves whether the risk indicator is derived from the performance of a specific technology. It is tempting to use the myriad of information available in the medical record, but this could potentially bias risk estimates across patients and providers. For instance, when possible, MedisGroups tries to minimize reliance on inherently unreliable physical examination findings by stipulating that information be drawn from the most "objective" source, such as technologically sophisticated diagnostic tests (for example, computed tomography, cystometrogram, endoscopy, cardiac catheterization). Although conceptually attractive, this strategy could result in biased comparisons across providers related to technological capabilities and practice patterns (Iezzoni and Moskowitz, 1988). Identical patients could receive different severity scores depending on the availability and use of a specific diagnostic test. Some severity measurement strategies take into account testing practices when choosing their clinical indicators (Keeler, Kahn, Draper, and others, 1990).

Despite the concerns outlined above, medical record information is nonetheless appealing in that it is the data that providers use daily to make diagnostic, therapeutic, and prognostic judgments. In using it for risk adjustment, however, one must recall that despite the attractions of detail and clinical credibility, medical record information may have limitations that could bias results.

Conclusion

Extensive interest in risk adjustment methods for health services research studies began only in the 1980s; therefore, in many ways the field is in its infancy. For example, relatively little work has been done to evaluate the performance of severity measures, even for predicting such comparatively easy-to-identify outcomes as costs of individual hospitalizations and in-hospital mortality. Only a handful of studies have yet compared performance across different measures: the few empirical, comparative studies that have been published are limited by small numbers of cases, nonrandom hospital samples, special patient populations (such as Medicare beneficiaries or specific conditions), or other limitations of the data bases employed (Alemi, Rice, and Hankins, 1990; Calore and Iezzoni, 1987; Charbonneau, Ostrowski, Poehner, and others, 1988; Thomas and Ashcraft, 1989, 1991). Therefore, much of the current information about the capabilities and performance of

severity systems comes from their developers and/or vendors. In addition, at the most basic level, researchers in the field are still struggling to develop appropriate statistical techniques for measuring performance of risk adjustment methods. Because a neat linear relationship may not exist between the level of risk and the occurrence of the outcome, risk adjustment raises many methodological challenges that have yet to be resolved.

Despite this relative paucity of empirical evidence, current research initiatives demand that choices be made: as stated at the outset, evaluations of medical effectiveness require some accounting for patient risk. The safest option in the absence of information is to choose a risk adjustment strategy that is most consistent conceptually with the methodological requirements of the research (for example, unit of analysis, timeframe, nature of the outcomes and treatments to be evaluated) and has the best clinical face validity given the overall research goals. When one thinks about face validity, it may be useful to consider each potential dimension of risk (see Table 2) and whether it is clinically pertinent to the specific research context.

One important choice is whether to use an existing measure or to develop a new approach specifically for the research study. Developing a risk adjustment method is a large, conceptually daunting, and potentially expensive undertaking, requiring extensive clinical input, pilot testing, and a credible means for "validation." Because of these challenges, many researchers opt to use an existing measure, even if it has certain conceptual or practical drawbacks. This alternative is reasonable and can often be made more palatable by adding supplemental information that addresses issues specific to the particular study. Often the available data will circumscribe the options for risk adjustment and will in large measure dictate the choice. When data limitations are weighed, it is important to delineate the potential sources of bias described above so the data can be used to their maximum advantage but with adequate caution.

Finally, as stated at the outset, even with the most detailed and accurate clinical data, it will not be possible to account fully for patient risk. Scenarios such as that of patient J in Table 3 are common, in which clinical judgment reliably predicts an inevitable outcome; only somewhat less common is the situation of patient I, who suffers an unanticipated outcome. Early in this century, Boston surgeon Ernest Amory Codman (1934) confronted a similar problem. He had developed the "end result idea ... the common-sense notion that every hospital should follow every patient it treats, long enough to determine whether or not the treatment has been successful, and then to inquire 'if not, why not' with a view to preventing similar failures in the future." He acknowledged the importance of adjusting for patient risk of bad outcomes, observing:

> For the man who practices surgery, there are two kinds of mortality—chance and intentional. Chance mortality is the kind which occurs unexpectedly, and which no amount of foresight can prevent. It is caused by unanticipated Calamities or Catastrophes. Death from pulmonary embolism is a good example ... Is it not possible to determine what this percentage of danger is, just as easily as it is to compute fire risk? ... Intentional mortality is incurred by the chief surgeon when he attempts cases in which the condition is acknowledged to be grave. It is speculative—like gambling against known chances in a game in which skill, judgment, and luck all count. (Codman, 1917)

Unlike Codman, we decades later have often forgotten the tremendous power of "random events." For example, one recent study found that 56-82 percent of excess hospital-level mortality for Medicare patients with acute myocardial infarction or congestive heart failure could result from random variation (Park, Brook, Kosecoff, and others, 1990). Given the constraints of uncertainty and the infinite variability of human biology, no method can be expected to provide a perfect prediction of patient risk. Thus, risk adjustment is a necessary but inherently imperfect tool.

References

Alemi, F., Rice, J., and Hankins, R. (1990). Predicting in-hospital survival of myocardial infarction. A comparative study of various severity measures. *Medical Care,* 28, 762-775.

American Cancer Society. (1989). *Cancer and the poor: A report to the nation.* Atlanta: Author.

American Hospital Association. (1979). *ICD-9-CM coding handbook for entry-level coders, with answers.* Chicago: American Hospital Publishing, Inc.

American Medical Association. (1992). *Current procedural terminology.* Chicago: Author.

Anderson, R.E., Hill, R.B., and Key, C.R. (1989). The sensitivity and specificity of clinical diagnostics during five decades. Toward an understanding of necessary fallibility. *Journal of the American Medical Association,* 261, 1610-1617.

Averill, R.F., McGuire, T.E., Manning, B.E., and others. (1989). *A study of the relationship between severity of illness and hospital cost in New Jersey.* New Haven, CT: Health Systems International.

Blumberg, M.S. (1991). Biased estimates of expected acute myocardial infarction mortality using MedisGroups admission severity groups. *Journal of the American Medical Association,* 265, 2965-2970.

Breslow, L. (1989). Health status measurement in the evaluation of health promotion. *Medical Care,* 27, S205-S216.

Brewster, A.C., Karlin, B.G., Hyde, L.A., and others. (1985). MEDISGRPS: A clinically based approach to classifying hospital patients at admission. *Inquiry,* 12, 377.

Brook, R.H., Ware, J.E., Davies-Avery, A., and others. (1979). Overview of adult health status measures fielded in RAND's health insurance study. *Medical Care,* 17(Suppl.), 1-131.

Burnum, J.F. (1989). The misinformation era: The fall of the medical record. *Annals of Internal Medicine*, 110, 482-484.

Calore, K.A. and Iezzoni, L.I. (1987). Disease staging and PMCs: Can they improve DRGs? *Medical Care*, 25, 724-735.

Charbonneau, D., Ostrowski, C., Poehner, E.T., and others. (1988). Validity and reliability issues in alternative patient classification systems. *Medical Care*, 26, 800-813.

Charlson, M.E., Pompei, P., Ales, K.L., and MacKenzie, C.R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases*, 40, 373-383.

Codman, E.A. (1917). *A study in hospital efficiency as demonstrated by the case report of the first five years of a private hospital.* Boston: Thomas Todd Company, Printers.

Codman, E.A. (1934). T*he shoulder. Rupture of the supraspinatus tendon and other lesions in or about the subacromial bursa* (pp. v-xi). Boston: Thomas Todd Company, Printers.

Conklin, J.E. and Houchens, R.L. (1987). *DRG refinement using measures of disease severity.* Santa Barbara, CA: SysteMetrics/McGraw-Hill.

Conklin, J.E., Lieberman, J.V., Barnes, C.A., and Louis, D.Z. (1984). Disease staging: Implications for hospital reimbursement and management. *Health Care Financing Review*, Ann. Suppl., 13-21.

Daley, J., Jencks, S., Draper, D., and others. (1988). Predicting hospital-associated mortality for Medicare patients: A method for patients with stroke, pneumonia, acute myocardial infarction, and congestive heart failure. *Journal of the American Medical Association,* 260, 3617-3624.

Detsky, A.S., Abrams, H.B., McLaughlin, J.R., and others. (1986). Predicting cardiac complications in patients undergoing non-cardiac surgery. *Journal of General Internal Medicine*, 1, 211-219.

Deyo, R.A. and Patrick, D.L. (1989). Barriers to the use of health status measures in clinical investigation, patient care, and policy research. *Medical Care*, 27, S254-S268.

Dubois, R.W., Rogers, W.H., Moxley, J.H., and others. (1987). Hospital inpatient mortality. Is it a predictor of quality? *New England Journal of Medicine*, 317, 1674-1680.

Eddy, D.M. (1984). Variations in physician practice: The role of uncertainty. *Health Affairs,* 3, 74-89.

Feigl, P., Glaefke, G., Ford, L., and others. (1988). Studying patterns of cancer care: How useful is the medical record? *American Journal of Public Health*, 78, 526-533.

Feinstein, A.R. (1988). ICD, POR, and DRG. Unsolved scientific problems in the nosology of clinical medicine. *Archives of Internal Medicine*, 148, 2269-2274.

Fetter, R., Freeman, J., Park, H., and others, and Health Systems Management Group (1989). *DRG refinement with diagnostic specific comorbidities and complications: A synthesis of current approaches to patient classification. Final Report* (prepared for the Health Care Financing Administration under Cooperative Agreement Nos. 15-C-98930/1-01 and 17-C-98930/1-0251). New Haven, CT: Health Systems Management Group.

Fetter, R.B., Shin, Y., Freeman, J.H., and others. (1980). Case mix definition by diagnosis related groups. *Medical Care,* 18(Suppl.), 1-53.

Gertman, P.M. and Lowenstein, S. (1984). A research paradigm for severity of illness: Issues for the diagnosis-related group system. *Health Care Financing Review*, Ann. Suppl., 79-90.

Gjorup, T., Bugge, P.M., and Jensen, A.M. (1984). Interobserver variation in assessment of respiratory signs. *Acta Medica Scandinavia*, 216, 61-66.

Goldman, D.R., Brown, F.H., Levy, W.K., and others (Eds.). (1982). *Medical care of the surgical patient.* Philadelphia: J.B. Lippincott.

Goldman, L. (1983). Cardiac risks and complications of noncardiac surgery. *Annals of Internal Medicine*, 98, 504-513.

Goldman, L., Caldera, D., Nussbaum, S.R., and others. (1977). Multifactorial index of cardiac risk in noncardiac surgical procedures. *New England Journal of Medicine*, 297, 845-850.

Gonnella, J.S. (Ed.). (1987). *Disease staging clinical criteria* (3rd ed.). Santa Barbara, CA: SysteMetrics/McGraw-Hill.

Gonnella, J.S., Hornbrook, M.C., and Louis, D.Z. (1984). Staging of disease. A case-mix measurement. *Journal of the American Medical Association*, 251, 637-644.

Gonnella, J.S., Louis, D.Z., and McCord, J.J. (1976). The staging concept—an approach to the assessment of outcome of ambulatory care. *Medical Care*, 14, 13-21.

Green, J., Passman, L.J., and Wintfeld, N. (1991). Analyzing hospital mortality. The consequences of diversity in patient mix. *Journal of the American Medical Association*, 265, 1849-1853.

Green, J., Wintfeld, N., Sharkey, P., and Passman, L.J. (1990). The importance of severity of illness in assessing hospital

mortality. *Journal of the American Medical Association*, 263, 241-246.

Greenfield, S., Aronow, H.U., Elashoff, R.M., and Wantanabe, D. (1988). Flaws in mortality data. The hazards of ignoring comorbid disease. *Journal of the American Medical Association*, 260, 2253-2255.

Greenfield, S., Blanco, D.M., Elashoff, R.M., and Ganz, P.A. (1987). Patterns of care related to age of breast cancer patients. *Journal of the American Medical Association*, 257, 2766-2770.

Horn, S.D., Bulkley, G., Sharkey, P.D., and others. (1985). Interhospital differences in severity of illness. Problems for prospective payment based on diagnosis-related groups (DRGs). *New England Journal of Medicine*, 313, 20-24.

Horn, S.D. and Horn, R.A. (1986). The computerized severity index. A new tool for case-mix management. *Journal of Medical Systems*, 10, 73-78.

Hsia, D.S., Krushat, M., Fagan, A.B., and others. (1988). Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. *New England Journal of Medicine*, 318, 352-355.

Iezzoni, L.I. (1989). Measuring the severity of illness and case mix. In Goldfield, N. and Nash, D.B. (Eds.), *Providing quality care: The challenge to clinicians* (pp. 70-105). Philadelphia: American College of Physicians.

Iezzoni, L.I., Ash, A.S., Cobb, J.B., and Moskowitz, M.A. (1988). Admission MedisGroups score and the cost of hospitalizations. *Medical Care*, 26, 1068.

Iezzoni, L.I., Ash, A.S., Coffman, G.A., and Moskowitz, M.A. (1991a). Admission and mid-stay MedisGroups scores as predictors of death within 30 days of hospital admission. *American Journal of Public Health*, 81, 74-78.

Iezzoni, L.I., Ash, A.S., Coffman, G.A., and Moskowitz, M.A. (1991b). Admission and mid-stay MedisGroups scores as predictors of hospitalization charges. *Medical Care*, 29, 210-220.

Iezzoni, L.I. and Moskowitz, M.A. (1988). A clinical assessment of MedisGroups. *Journal of the American Medical Association*, 260, 3159.

Iezzoni, L.I., Moskowitz, M.A., and Ash, A.S. (1988). *The ability of MedisGroups and its clinical variables to predict cost and in-hospital death* (prepared for the Health Care Financing Administration under Cooperative Agreement No. 18-C-98526/1-04). Waltham, MA: Brandeis University, Health Policy Research Consortium.

Iezzoni, L.I., Moskowitz, M.A., and Daley, J. (1989). *A description and clinical assessment of the computerized severity index* (prepared for the Health Care Financing Administration under Cooperative Agreement No. 18-C-98526/1-05). Waltham, MA: Brandeis University, Health Policy Research Consortium.

Institute of Medicine. (1977). *Reliability of hospital discharge abstracts.* Washington: National Academy of Sciences.

Jencks, S.F., Williams, D.K., and Kay, T.L. (1988). Assessing hospital-associated deaths from discharge data. The role of length of stay and comorbidities. *Journal of the American Medical Association*, 260, 2240-2246.

Jenkins, C.D. (1983). Social environment and cancer mortality in men. *New England Journal of Medicine*, 308, 395-398.

Jewell, E.R. and Persson, A.V. (1985). Preoperative evaluation of the high-risk patient. *Surgical Clinics of North America*, 65, 3-19.

Johnson, J.E. and Carpenter, J.L. (1986). Medical house staff performance in physical examination. *Archives of Internal Medicine*, 146, 937-941.

Kaplan, S.H. and Ware, J.E. (1989). The patient's role in health care and quality assessment. In Goldfield, N. and Nash, D.B. (Eds.), *Providing quality care. The challenge to clinicians* (pp. 25-68). Philadelphia: American College of Physicians.

Keeler, E.B., Kahn, K.L., Draper, D., and others. (1990). Changes in sickness at admission following the introduction of the prospective payment system. *Journal of the American Medical Association*, 264, 1962-1968.

Knaus, W.A., Draper, E.A., Wagner, D.P., and Zimmerman, J.E. (1985). APACHE II: A severity of disease classification system. *Critical Care Medicine*, 13, 818-829.

Knaus, W.A., Draper, E.A., Wagner, D.P., and Zimmerman, J.E. (1986). An evaluation of outcome from intensive care in major medical centers. *Annals of Internal Medicine*, 104, 410-418.

Kosecoff, J., Fink, A., Brook, R.H., and Chassin, M.R. (1987). The appropriateness of using a medical procedure. Is information in the medical record valid? *Medical Care*, 25, 196-201.

Lloyd, S.S. and Rissing, J.P. (1985). Physician and coding errors in patient records. *Journal of the American Medical Association*, 254, 1330-1336.

Lohr, K.N. (Ed.). (1989). Advances in health status assessment: Conference proceedings. *Medical Care*, 27, S1-S294.

Manton, K.G., Patrick, C.H., and Johnson, K.W. (1987). Health differentials between blacks and whites: Recent trends in mortality and morbidity. *Milbank Quarterly*, 65, 129-199.

McMahon, L.F. and Smits, H.L. (1986). Can Medicare prospective payment survive the ICD-9-CM disease classification system? *Annals of Internal Medicine*, 104, 562-566.

Meinert, C.L. (1986). *Clinical trials. Design, conduct, and analysis.* New York: Oxford University Press.

Mendenhall, S. (1984). DRGs must be changed to take patient's illness severity into account. *Modern Healthcare*, 14, 86-88.

Otten, M.W., Teutsch, S.M., Williamson, D.F., and Marks, J.S. (1990). The effect of known risk factors on the excess mortality of black adults in the United States. *Journal of the American Medical Association*, 263, 845-850.

Park, R.E., Brook, R.H., Kosecoff, J., and others. (1990). Explaining variations in hospital death rates, randomness, severity of illness, quality of care. *Journal of the American Medical Association*, 264, 484-490.

Pollack, M.M., Ruttimann, U.E., and Getson, P.R. (1988). Pediatric Risk of Mortality (PRISM) score. *Critical Care Medicine*, 16, 1110-1116.

Prospective Payment Assessment Commission. (1990). *Report and recommendations to the Secretary, U.S. Department of Health and Human Services*. Washington: U.S. Government Printing Office.

Prospective Payment Assessment Commission. (1991). *Medicare and the American health care system. Report to the Congress* (p. 49). Washington: U.S. Government Printing Office.

Raftery, E.B. and Holland, W.W. (1967). Examination of the heart: An investigation into variation. *American Journal of Epidemiology*, 85, 438-444.

Romm, F.J. and Putnam, S.M. (1981). The validity of the medical record. *Medical Care*, 19, 310-315.

Savino, J.A. and Del Guercio, L.R.M. (1985). Preoperative assessment of high-risk surgical patients. *Surgical Clinics of North America*, 65, 763-791.

Schneider, A.J.L. (1985). Assessment of risk factors and surgical outcome. *Surgical Clinics of North America*, 65, 1113-1126.

Simborg, D.W. (1981). DRG creep: A new hospital-acquired disease. *New England Journal of Medicine*, 304, 1602-1604.

Slee, V.N. (1978). The International Classification of Diseases: 9th Revision (ICD-9). *Annals of Internal Medicine*, 88, 424-426.

Spilker, B., Molinek, F.R., Johnston, K.A., and others. (1990). Quality of life bibliography and indexes. *Medical Care*, 28(Suppl), DS1-DS77.

Starfield, B., Steinwachs, D., Morris, I., and others. (1979). Concordance between medical records and observations regarding information on coordination of care. *Medical Care*, 17, 758-766.

Steinwachs, D.M. (1989). Application of health status assessment measures in policy research. *Medical Care*, 27, S12-S26.

Steinwald, B. and Dummit, L.A. (1989). Hospital case-mix change: Sicker patients or DRG creep? *Health Affairs*, 8, 35-47.

Stewart, A.L., Greenfield, S., Hays, R.D., and others. (1989). Functional status and well-being of patients with chronic conditions. Results from the medical outcomes study. *Journal of the American Medical Association*, 262, 907-913.

SysteMetrics. (1988). *Disease staging Q-scale*. Santa Barbara, CA: SysteMetrics/McGraw-Hill.

Thomas, J.W. and Ashcraft, M.L.F. (1989). Measuring severity of illness: A comparison of interrater reliability among severity methodologies. *Inquiry*, 26, 483-492.

Thomas, J.W. and Ashcraft, M.L.F. (1991). Measuring severity of illness: Six severity systems and their ability to explain cost variations. *Inquiry*, 28, 39-55.

U.S. Congress, Office of Technology Assessment. (1985). *Medicare's prospective payment system: Strategies for evaluating cost, quality, and medical technology* (OTA-H-262, pp. 195-199). Washington: U.S. Government Printing Office.

Vibbert, S. (1991). UCDS: HCFA's ambitious automated review project. *Medicine and Health: Perspectives*. 45(11), insert.

Vladeck, B.C. (1984). Medicare hospital payment by diagnosis-related groups. *Annals of Internal Medicine*, 100, 576-591.

Wachter, R.M., Luce, J.M., Hearst, N., and Lo, B. (1989). Decisions about resuscitation: Inequities among patients with different diseases but similar prognoses. *Annals of Internal Medicine*, 111, 525-532.

Wagner, D.P. and Draper, E.A. (1984). Acute physiology and chronic health evaluation (APACHE II) and Medicare reimbursement. *Health Care Financing Review*, Ann. Suppl., 91-105.

Wagner, D.P., Knaus, W.A., and Draper, E.A. (1986). Physiologic abnormalities and outcome from acute disease. Evidence for a predictable relationship. *Archives of Internal Medicine*, 146, 1389-1396.

Whiting-O'Keefe, Q.E., Henke, C., and Simborg, D.W. (1984). Choosing the correct unit of analysis in medical care experiments. *Medical Care*, 22, 1101-1114.

Wiener, S. and Nathanson, M. (1976). Physical examination. Frequently observed errors. *Journal of the American Medical Association*, 236, 852-855.

Young, W.W. (1984). Incorporating severity of illness and comorbidity in case-mix measurement. *Health Care Financing Review*, Ann. Suppl., 23-31.

Young, W.W, Swinkola, R.B., and Zorn, D.M. (1982). The measurement of hospital case mix. *Medical Care*, 20, 501-512.

Zimmerman, J.E., Knaus, W.A., Sharpe, S.M., and others. (1986). The use and implications of do not resuscitate orders in intensive care units. *Journal of the American Medical Association*, 255, 351-356.

Risk Adjustment for Medical Outcome Studies: Are We Opening Pandora's Box?

Constantine A. Gatsonis, Ph.D.

Risk Adjustment and Randomization

A key component of the recent growth in outcomes research has been the proliferation of proposals for "risk adjustment" strategies. The importance of the use of such strategies is increasing, particularly as outcomes research strives to expand its focus by moving beyond the level of description and into the exploration of causal links. It would not be an exaggeration to say that the ultimate success or failure of this expansion will depend, to a large extent, on the development of scientifically credible and practically feasible risk adjustment strategies.

A detailed discussion of the technical aspects of planning for and implementing risk adjustment would inevitably touch on large areas of statistics and epidemiology. Iezzoni (in this volume) quite properly leaves the technical discussion aside and concentrates on the broader conceptual and methodological issues. She provides a substantial and incisive review of the state of the art in risk adjustment, which will be valuable to a broad spectrum of researchers. The review can serve as a thorough introduction to the subject for beginners, but it also has much to offer those of us who have already grappled with the thorny issues in this area.

It seems appropriate to begin this commentary at the point where Iezzoni's discussion ends: "Thus, risk adjustment is a necessary but inherently imperfect tool." Since the necessity of risk adjustment is no longer an issue for debate and the imperfection of the tools for carrying it out becomes quickly transparent to anyone who has worked with them, it would appear that the important message in the closing sentence is that risk adjustment is "inherently imperfect." To be sure, one could say the same thing about almost any method or tool in the applied sciences, but the problems with risk adjustment seem to be quite serious indeed. In a very real sense, the whole endeavor amounts to opening Pandora's box, which contains the multitude of patient characteristics and other factors that may influence therapeutic response. This box was traditionally kept closed through the use of randomization in order to provide a balance of risk factors across experimental groups in the context of well-designed prospective studies.

The opening of Pandora's box is often not a matter of choice. It is simply not practically possible to carry out randomized clinical trials (RCTs) to cover all categories of treatments, procedures, patient subgroups, disease subtypes, and so on. Thus, the need to carry out nonrandomized studies is pressing, but the conduct of such studies should be seen as complementary to RCTs rather than as a replacement. Whatever one's hopes for the distant future may be, the present state of methodological development of observational outcome studies and the nature of the obstacles encountered thus far offer little justification for expecting that such studies will obviate the need for RCTs in the foreseeable future. The optimistic view of this tension between the two types of studies is that nonrandomized outcome studies will provide a wealth of new information to be combined with the results from RCTs in the process of making recommendations for clinical and policy decisions.

Iezzoni's paper (in this volume) provides a useful road map of conceptual issues and challenges that are encountered in the formulation and implementation of risk adjustment strategies. This commentary concentrates on the more technical aspects of design and analysis for such strategies. The discussion is divided into two parts. First, risk adjustment is formulated as part of the process of developing statistical models of the link between process and outcome. Such a model building activity is always carried out for a *particular* clinical condition(s), using a specific data base available for this purpose. This conceptualization implies that a risk adjustment strategy can proceed successfully without either requiring the use of or leading to the development of multipurpose "indices" for constructs such as severity or comor-

Dr. Gatsonis is Assistant Professor of Biostatistics, Department of Health Care Policy, Harvard Medical School.

bidity, which are applicable across studies or conditions. In fact, the uncritical use of such indices may confuse rather than facilitate matters in many studies. Second, a review of methodological and statistical issues is used to support the argument that the most immediate gains in risk adjustment strategies can be achieved by improvements in experimental design and data quality. Although further methodological development may be necessary in several instances, most analytic issues are not unique to risk adjustment and can be addressed by adapting already developed statistical approaches.

Risk Adjustment as Statistical Model Building

The inherent imperfection of risk adjustment stems from the fundamental realization that the phenomena under study are so highly complex and multidimensional that a complete account of all the factors influencing the outcome would require an enormous amount of data and a very complex statistical model to describe them. In any particular situation, a series of judgment calls is needed to make the conduct of the study practically feasible.

Iezzoni's review (in this volume) provides a comprehensive list of the categories of factors that might be included in a risk adjustment strategy. Although it is relatively easy to provide verbal descriptions of the various categories, such as "clinical stability," "severity," and "comorbidity," the quantification of these concepts is far from easy or uncontroversial. The immense array of variables describing the condition of the patient at a given time period needs to be made manageable. This is done as part of the process of fitting statistical models to the data that describe the relation of process to outcome.

The overriding aim in risk adjustment is to find a parsimonious representation of those patient characteristics that are most likely to have a strong relation to the specific outcome and that could, without proper attention, confound the results of the particular analysis. Such a representation is constructed as part of the process of statistical model building for the specific data at hand. This process will often involve the use of techniques for data compression and will be guided by both statistical/methodological considerations and clinical input. If specific indices of severity or comorbidity, for instance, are available for a condition under study, the investigator may benefit from their use. For example, researchers in stroke or myocardial infarction who have access to a rich clinical data base may use the indices incorporated in the Medicare Mortality Predictor System (Jencks, Daley, Draper, and others, 1988), but medical abstract data cannot be used for research on the same conditions. However, a perfectly adequate risk adjustment strategy need not—and most likely will not—culminate in the derivation of stand-alone indices for constructs such as severity or comorbidity, which can then be used independently in other contexts, such as making clinical or reimbursement decisions. In fact, it is fair to say that the usefulness of an index in a specific clinical context is often inversely related to how broadly relevant the index was constructed to be. It is not a surprise that many of the quests for multipurpose indices have been followed by a host of often polemical rejoinders, which describe the limitations of the index in specific clinical settings and for particular uses (for a recent example, see the discussion in Blumberg, 1991).

It seems then that we would be well advised to reexamine the need for use of indices in outcomes research and, if possible, to decouple it from the process of risk adjustment. The construction of data-based instruments that can be used to predict clinical course and outcome at the individual patient level is certainly a most worthwhile research goal, but it should be seen as distinct from that of providing an adequate accounting of the main factors and sources of variability in a model relating process to outcome.

Statistical Issues

The statistical issues in risk adjustment are numerous and multifaceted, although not necessarily unique to this area of empirical research. Issues, such as data reduction, model selection, checking for linearity of effects and the presence of interactions, detecting outliers, examining residuals, and assessing model fit, are generic problems in statistical model building. There exists by now a considerable body of knowledge and experience that can be brought to bear on these matters, although the specific applications may not always be routine.

One category of problems, which feature prominently in risk adjustment and deserve special attention, relates to the quality and availability of the data. In particular:

1. Numerous data elements may be missing, often not at random.
2. Data can be selectively available, thus making it difficult to assess reliably the presence or absence of a specific condition or medical finding.
3. Data coding errors may be pervasive.
4. Even when data are present and properly recorded, they may contain factual errors, reflecting, for instance, judgments made by diagnosticians and clinicians.

A discussion of the above problems should be prefaced by the observation that attention to experimental design and data collection procedures can play a major role in minimizing the extent and impact of such difficulties.

This theme will be expanded below in the context of specific proposals for improvements. Partly as a motivation for these proposals, a brief review of methodological aspects of the data quality and availability issues is in order.

With regard to problems of missing data, much progress has been made in the statistical literature in recent years (see Little and Rubin, 1987, for a comprehensive review). As a result of these developments, missing data problems can now be addressed with model-based techniques and multiple imputation (Little and Rubin, 1987; Rubin, 1987). These approaches, however, require the use of additional information about the pattern of "missingness," which may be difficult to obtain in practice. Their success is also crucially dependent on how extensive the missing data are. From a practical point of view, they should be seen as a last resort effort, after all attempts to locate the missing data elements have been made during the design and data collection phase of the study.

A related issue to missing data is the problem of selective availability of some data elements, which is almost invariably present in the clinical information contained in administrative data bases. This problem can become the source of considerable confusion in any risk adjustment effort by introducing artifactual effects and thus confounding the investigation. Consider, for example, the finding reported by some investigators that the presence of several comorbid conditions (such as arrhythmia or hypertension) appears to have a protective effect in patients who have suffered an acute myocardial infarction. Although it might be possible to construct plausible clinical explanations for some of these findings, the investigator needs to address the possibility that selective use of coding may have produced such apparently counter-intuitive phenomena.

The problems of missing or selectively available data are further compounded by the frequent presence of coding and measurement errors. Coding errors have been shown to be extensive even for key data items such as the primary diagnosis (Hsia, Krushat, Fagan, and others, 1988; Iezzoni, Burnside, Sickles, and others, 1988). Measurement errors, although undoubtedly present, have not been given any attention in the risk adjustment literature to date. A related class of data errors, namely those arising from factual mistakes made by clinicians and diagnosticians, also remains uncharted territory. Some statistical methodology is already in place for addressing specific problems of misclassification and measurement error (see Chen, 1989, for a recent review of methods for misclassification, and Fuller, 1987, for errors in variables). The applicability of the existing methodology to problems in risk adjustment is an open question at this point, since we lack both the auxiliary information required to implement such procedures and the experience of having attempted to use them. Further development of methodology may be needed once the magnitude and nature of the data error problems have been identified. For example, errors-in-variables methods tend to rely on severe assumptions on the structure of the error (such as normality or at least symmetry of the error distribution), which rarely will be applicable to the analysis of medical outcomes data. As another example, misclassification has been studied in the context of contingency tables analysis, but it seems that considerable progress is needed in the study of its effect in logistic regression modeling.

Experimental Design and Data Collection: Major Areas for Improvement

The discussion of the analytic and data quality issues identified several areas where statistical "technology transfer" is possible as well as some areas that may require new development. This discussion, however, should also have made it quite evident that there is much to be gained by focusing attention on the design and the data collection process of the outcome studies that will eventually require the implementation of risk adjustment. In fact, it would not be an exaggeration to say that, at this point in the development of medical outcome studies, improvements in study design and data collection constitute the single most important way to make substantial progress. To put it in another way, before we spend a lot of time developing statistical "fixes" for problems with the data, it would be well worth our effort to take care of some of these problems as we design and implement the study. Below is a list of some of the major aspects of design and data collection where improvements can be made.

1. Many risk adjustment problems stem from the fundamental reality that often outcome studies are retrospective investigations, utilizing data bases that were not originally designed for this purpose. It is not difficult to see then that there is much to be gained by designing and implementing prospective studies of medical outcomes. Prospective designs will make it possible to exercise much more effective control over the multitude of factors affecting outcome than is presently the case in retrospective analyses of existing data bases. To be sure, prospective studies require considerably more effort and expense than retrospective ones; however, it seems that such resource intensity is necessary if we expect to bring the results from nonrandomized outcome studies to a

level of scientific sophistication similar to that found in RCTs.

2. Both prospective and retrospective studies will benefit from having a careful definition of the clinical cohort of patients under study. The broadness of the study population will always have to be determined on the basis of the strength of the available data. One of the often cited strengths of observational outcomes studies when compared with RCTs is the ability of the former to provide information that is applicable to a much broader spectrum of patients than is possible for RCTs. The other side of this coin, however, is that generalizability is often inversely proportional to the depth and reliability of the conclusions. To the extent that risk adjustment is called upon to rectify some of the problems caused by the lack of randomization, clarity (and realism) in the definition of the clinical cohort(s) will provide a reliable platform on which an outcome study can be built.

3. The quality and validity of the data will have to be assessed in both prospective and retrospective studies (especially the latter). Auxiliary studies may often be necessary in order to assess the accuracy of the data, but unfortunately, such studies are not carried out routinely. In fact, we are all too familiar with the tendency in outcomes research to think that our problems will be solved simply by the sheer magnitude of our data bases. The advent of modern statistics has provided convincing evidence that complete enumerations are not needed in order to arrive at accurate answers for most research and policy questions. What is absolutely needed, however, are quality and accuracy of the obtained data elements. The situation brings to mind an analogy from hydrostatics. A dam built on a lake covering an area equal in size to Minnesota is going to be useless if the depth of the lake is only a couple of feet! We need depth in our data bases, and the quality and accuracy of the data are critical components of such depth.

Improvements in the three areas of design and data collection outlined above will diminish the effects of major sources of "noise" in the outcome data, and thus, they will greatly enhance our ability to carry out risk adjustment strategies successfully. To be sure, these improvements are not going to be panaceas. Many analytic problems will remain, and they will require innovative statistical approaches; but major progress will have been accomplished. If we are going to open Pandora's box, and it seems that we are well along in the process of doing that, we should be prepared to face the challenges.

References

Blumberg, M.S. (1991). Biased estimates of expected acute myocardial infarction mortality, using MedisGroups admission severity groups. *Journal of the American Medical Association,* 265, 2965-2970.

Chen, T.T. (1989). A review of methods for misclassified categorical data in epidemiology. *Statistics in Medicine*, 8, 1095-1106.

Fuller, W. (1987). *Measurement error models.* New York: John Wiley & Sons, Inc.

Hsia, D.C., Krushat, W.M., Fagan, A.B., and others. (1988). Accuracy of diagnostic coding for Medicare patients under the prospective payment system. *New England Journal of Medicine*, 318, 352-355.

Iezzoni, L.I., Burnside, S., Sickles, L., and others. (1988). Coding of acute myocardial infarction. *Annals of Internal Medicine*, 109, 745-751.

Jencks, S.F., Daley, J., Draper, D., and others. (1988). Interpreting hospital mortality data: The role of clinical risk adjustment. *Journal of the American Medical Association*, 260, 3611-3616.

Little, R. and Rubin, D. (1987). *Statistical analysis with missing data.* New York: John Wiley & Sons, Inc.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys.* New York: John Wiley & Sons, Inc.

The Relationship Between Use of Health Care Services and Health Status: Dilemmas in Measuring Medical Outcome in Low-Income and Racial/Ethnic Populations

Ruth E. Zambrana, Ph.D.

Low-income and racial/ethnic populations constitute between one-quarter and one-third of the U.S. population. The consensus in the field is that poverty or low-income status is the strongest predictor of both the use of health services and health outcome. Yet there is limited understanding of the complex processes and factors that influence the pathways to unfavorable health outcome. Delayed use of health care, organizational and financial barriers, and knowledge, attitudes, and beliefs have been defined as a set of interrelated factors that influence outcomes. It has become painfully clear, however, that these factors do not provide sufficient explanatory power to enable us to understand patient outcomes following medical care. Recent evidence suggests that the right questions are not being asked, and that there is a need to improve our conceptual paradigms, methodological approaches, and data collection procedures (Cramer, 1987; Eberstein, 1989; Eberstein, Nam, and Hummer, 1990). In a recent National Center for Health Statistics (NCHS) publication, it is acknowledged that "the quality of the data is going to be no better than the quality of the questions asked" (Sirken, Bercini, and Jobe, 1990).

The purposes of this paper are to: examine the scientific approaches in the study of special populations; discuss a set of factors that require a scientific consensus—that is, the formulation of operational definitions that are derived from a theoretical and conceptual knowledge base; present methodological caveats that may hamper the identification of special populations and the collection of medical outcome data on low-income and racial/ethnic populations; propose a set of research principles that can guide the development of cross-cultural methodologies for special or culturally and racially distinct populations; and provide suggestions to improve data quality. The data for this paper are derived from two principal sources: one, a review of the substantive literature in the field of methodological approaches for low-income and special populations, and two, research experiences and observations during the last 6 years in the conduct of a project entitled "Mediators of Birth Outcome in Low-Income Racial-Ethnic Women in Los Angeles." More specifically, methodological issues in identifying and matching medical record files and identifiers for locating respondents, with a particular focus on Hispanic populations, are central to the discussion.

Dr. Zambrana is Associate Professor, School of Social Welfare, University of California, Los Angeles.

Scientific Approaches to Special Populations

The scientific approach to understanding the complex needs of individuals, particularly in the field of health, has been driven by a rigid, narrow, and homogeneous set of principles that has not permitted the exploration of diversity in the health trajectories of individuals. The fields of health services research, biomedical research, demography, and social epidemiology have narrowly and inconsistently defined the parameters of disease patterns as unidimensional—that is, occurring outside the context of a social environment, a psychosocial context, and a family context. The developmental life of a disease and its medical outcome cannot be measured in a vacuum. Or perhaps they can, but the data provide little knowledge regarding the contributory factors and thus fail to assist in developing future interventions that will influence the course of the medical outcome.

In the last 6 years as I encountered problems in my own research for which I did not readily have answers, I carefully reviewed several research design books in the field to obtain assistance. Recent books in the field (Aday, 1989; Doherty and Campell, 1988; Grady and Wallston,

1988), with the exception of one (Marin and Marin, 1991), have contained limited or no recognition of special research concerns in studies of racial/ethnic populations, such as cross-cultural applicability of item or scale selection in instrument development. Social scientists in the fields of public health (Neresian, 1988) and demography (Cramer, 1987; Rogers, 1989), however, have forged ahead to document the obvious—that is, health outcome is partly a function of social inequality. In examining social differentials in infant mortality, Eberstein (1989, pp. 410-411) proposes two hypotheses:

1. Inequality in death rates is understood to reflect inequality in life conditions; and

2. The most important inequalities are considered fundamentally social.

While the first of these hypotheses predisposes one to consider the meaning of inequality and its mortality implications, the second focuses attention on social and economic, as opposed to genetic (e.g., racial) or medical (i.e., different diseases or disease processes) inequalities as the explanatory variables of interest.

Given the interest in testing these hypotheses, the next question is, "What are the indicators to assess the parameters of social inequality?" The lack of attention given to the importance of fundamental background population characteristics such as race, ethnicity, and level of maternal education (a well-established indicator of social class) is perhaps most readily observable in my recent review which aimed to identify racial and ethnic differences in medical risk disease patterns in pregnant women. For example, in the review by Cousins (1987) of pregnancy complications in diabetic women from 1965 to 1985, there was no mention of low-income or racial/ethnic differences or any suggestion in the agenda for future research to examine these factors as differentially influencing the rate of maternal complications and mortality. This finding was surprising in view of the extant knowledge in the field of maternal and child health (for example, that diabetes— especially gestational diabetes—and hypertension are more prevalent in Hispanic and black women) (Department of Health and Human Services [DHHS], 1986).

In a computer search covering the last 10 years which we conducted for our project on ethnic differences in medical risk in pregnancy, only 24 articles out of over 1,000 identified major ethnic and socioeconomic indicators for the sample. The lack of racial and ethnic identifiers in medical risk data may hinder our understanding of the pathways that influence low birth weight in poor and/or racial/ethnic women.

Another issue is related to the uses of our foundations of knowledge and the continued development of broader and more insightful explications of the factors that influence medical outcome and medical effectiveness. For instance, although the Institute of Medicine (1988) has clearly stated that there is no need for further research on barriers to prenatal care, the Robert Wood Johnson Foundation has funded additional studies to assess areas for which we seem to have identified the scientific knowledge base to implement appropriate programs and policies. For example, "Does removing financial barriers to prenatal care improve pregnancy outcomes for low income women?" "Will a significant increase in Medicaid payment for deliveries augment physician participation in the Medicaid program and improve access to care? How does the health status and the health care utilization of the uninsured differ from the insured?" (Health Care Financing and Organization, 1991, pp. 8-9) I wonder if there would be a lack of agreement that the responses to all these questions are in the significantly affirmative range. Lastly, the fact that the analysis of the Hispanic Health and Nutrition Examination Survey (H-HANES) data, which was collected in 1982-1984, was recently published (Trevino, 1990; Becerra, Hogue, Atrash, and Perez, 1991; Stroup-Benham and Trevino, 1991; Mendoza, Ventura Valdez, and others, 1991; Trevino, Moyer, Valdez, and Stroup-Benham, 1991; Ginzberg, 1991; Warner, 1991; American Medical Association, 1991; Hispanic Health, 1991; Health Status, 1991) and reflects the historical and contemporary avoidance of recognizing the needs of special populations in the United States.

In my view, the insistence of the health services research and social science community to persist in asking the same questions and confirm the obvious is not only disheartening but continues to erode the limited knowledge base acquired in the last decade. Repeatedly, I have asked myself the question, "Why do researchers persist in not recognizing the influence of social inequality?" These issues, I believe, must be viewed in the context of the norms that guide the research process in the United States.

At this juncture, there is a need to revisit the principles of the scientific method. In essence, the question is, "Have research scientists violated the basic principles of the scientific method?" That is:

1. Empirical research must be based on observation of the real world. In the real world, access to benefits is related to health outcome. The steady drop in infant mortality between 1965 and 1980 is the most notable example of Medicaid's contribution to improved infant outcome (Schorr, 1989, p. 125); and participa-

tion in the Department of Agriculture's Women, Infants, and Children (WIC) Program reduces premature births in high-risk mothers by 15-25 percent (Schorr, 1989, p. 129).

2. Systematic principles require that investigators report all data and procedures. I interpret this to include sample characteristics such as race, ethnicity, and socioeconomic position; as noted, few studies record the characteristics of the sample in their reporting procedures.

3. All scientific conclusions are deemed to be provisional and can be challenged, since we believe as scientists that knowledge is continuously evolving and that "scientific revolutions" continuously change our understanding of our real-world observations; yet we have not built on prior knowledge.

4. Objectivity, which mandates that investigators "attempt" to remove biases, is central to the research endeavor.

Thus, past and current research endeavors exhibit several features: one, they are generally individualistic—that is, guided by a principal investigator; two, limited value is placed on the role of collaboration either across disciplines or with the community or subjects; and three, the experiences of many researchers are not practice or community based. Perhaps Sultz (1991) has captured the essence of the issue when he states:

> As a nation, we have an outstanding capacity for avoiding the truth about matters of public health and public good and it can be attributed in large part to our historical commitment to individual freedom. (pp. 418-419)

Thus, there may be an incompatibility between individual freedom and social responsibility in research. The ethical questions that are raised are related to our responsibility to the society in which we live and the functions and uses of the research product. In December 1990, the National Institutes of Health/Alcohol, Drug Abuse, and Mental Health Administration issued a priority announcement that, "All those whose proposed research question would be answered inadequately if women or members of minority groups were not part of the research sample are advised to use such subjects or offer a compelling justification for their exclusion."

Johnson (1991) raises several bothersome questions, "Should scientists not have recognized those responsibilities without being forced to do so? Will they pay sufficient attention to other such justifiable social demands on their ability to produce reliable knowledge?" Perhaps some would argue that all this is too difficult and requires extensive time. On the other hand, one could also argue that bad research takes the same amount of time, the same number of subjects, the same human protection procedures, and the same number of copies for Washington as good systematic research.

The point to be made here is that cross-cultural or cross-ethnic research methodology is not a new invention but rather represents adherence to rigorous systematic research principles— that is, a sensitivity to and a real-world knowledge of the problems experienced by specific sectors of the society which are, in turn, adequately measured through a set of scientific and systematic steps. Most importantly, it recognizes the influence of poverty and its correlates as major determinants of health status, use of health care services, and medical outcome. Thus, the scientific principles to date have been loose—that is, guided by individual interpretation as opposed to a consensus by the scientific community. There have been few guidelines and operational definitions to guide the research endeavor, seriously impeding comparability across studies and across national data sets. If we are truly to improve our medical effectiveness outcome data base, we need to arrive at a scientific consensus on who and what we are studying and which variables we are to measure.

Scientific Consensus

Recent studies have identified the need to distinguish and measure more clearly those variables that can serve as accurate descriptors of low-income or vulnerable populations. Past studies have tended to ignore class distinctions among ethnic groups and nativity within ethnic groups. Thus, the data generated have not been representative of low-income groups and have not permitted comparisons across or within groups. The resistance to distinguishing between classes is related to many researchers' lack of experience and familiarity with the context of the lives of poor and racial and ethnic populations. There is also a tendency to perceive all poor and minority individuals as a homogeneous group—that is, the poor are minorities, and they have distinguishing characteristics that form a culture of poverty.

This dominant view is best exemplified by the current emphasis on the urban disadvantaged. We see two major strands in the mainstream literature. One, which is most applicable to Hispanics, blames the culture for its deficits. The second, a poverty deficit model, most applicable to the black population, focuses on the persistent concentration of poverty and transmission of poverty cycles.

For blacks and Hispanics, poverty has been a reality for decades and consistently related to lower use of services and higher morbidity and mortality patterns. Re-

cent data show that the percentage of poverty in these groups has increased. For example, in Los Angeles, Hispanic poverty rates increased from 22.4 percent in 1977 to 28.2 percent in 1987. For black women, increases in poverty and cuts in benefits have contributed, perhaps significantly, to increases in infant mortality, which has been referred to as a "capital crime" (Boone, 1989). Therefore, the issue is not to continue to relabel the phenomenon of poverty, but rather to recognize the different levels of socioeconomic position in the United States, measure class to assure comparability across and within groups, and recognize the powerful effects of class as it interacts with culture, ethnicity, nativity, and race.

Recent literature has documented differences within Hispanic groups and differences between black and Hispanic populations in pregnancy outcomes, especially infant mortality and low birth weight. One plausible explanation proposed for cross-ethnic differences is the buffering effects of culture. The cultural role has been broadly and inconsistently defined as the use of more preventive or healthy lifestyle practices or presence of positive attitudes toward pregnancy. Such practices include abstinence from alcohol, drugs, and cigarettes; use of self-care practices, such as planned pregnancy, because of a high value on motherhood and pregnancy; and the availability of good social support networks, including a higher likelihood of living with a male partner. For Mexican immigrant women, there may also be a self-selection factor operating, since the infant death rate in Mexico is 47 per 1,000 live births nationally, 20 per 1,000 live births in border towns (Warner, 1991), and 5.6 per 1,000 live births in the United States. For Puerto Rican women, however, this cultural buffering effect does not seem to hold. Infant mortality is 11.6 per 1,000 live births in the United States and 14.9 per 1,000 live births in Puerto Rico (Becerra, Hogue, Atrash, and Perez, 1991). Furthermore, for Mexican-American women, the cultural buffering factors may be eroding as a function of greater integration into this society, which is defined as acculturation (for example, use of alcohol or drugs).

The concept of acculturation is measured in many diverse ways (Cuellar, Harris, and Jasso, 1980; Olmedo and Padilla, 1978). The core factors include number of years in the United States, ability to speak English proficiently, educational level, and ethnic self-identification. The major finding in the acculturation literature is that acculturation is partly an indicator of socioeconomic position—that is, the more accultured tend to have higher levels of education and are more likely to use health services (Chesney, Chavira, Hall, and Gary, 1982; Deyo, Diehl, Hazuda, and Stern, 1985). These intra-ethnic differences also highlight the importance of distinguishing between Hispanic subgroups (Hayes-Bautista, 1990a, 1990b; Hayes-Bautista, Hurtado, and Valdez, 1990). It has been relatively well-established that Cubans and other Hispanic subgroups (the more educated Central and South Americans) approximate most closely the educational level of the white majority (Ginzberg, 1991). Thus, these differences are important to measure as individual attributes or background variables that are related to and influence health status, use of services, and medical outcome.

Areas of Consensus: Moving Toward a Conceptual Paradigm

At present, there seems to be sufficient scientific frustration in the social science community (in relation to the lack of comparability in similar studies and to contradictory and inconclusive findings) to warrant a consensus on what needs to be examined to move the field forward. I propose a series of indicators that would have a significant relationship to medical outcome. The challenge in the field frequently seems to be related to, "Why this indicator over that one?" and "Why do we need it?"

Race. Race is viewed as both a biological and a social indicator. In the United States, it is inextricably linked with socioeconomic position, ethnicity, and mortality and morbidity patterns (Byrd, 1990).

Ethnicity. Ethnicity is assuming a more prominent position in health services research with the unprecedented growth and immigration of diverse ethnic and racial groups to the United States—for example, Hmong (Vietnamese tribe), Filipinos, and Central Americans, especially those from El Salvador and Guatemala. Thus, the measurement of ethnicity requires an expansion to encompass more than a glimpse of a cultural group and its practices; it requires more precise data on place of birth (nativity), numbers of years in the United States, and indigenous language. For example, it cannot be assumed that Spanish is the primary language of all individuals from Spanish-speaking countries. In fact, Spanish is not the indigenous or primary language of the people from certain parts of El Salvador and certain southern regions of Mexico; it is the second language. This is an important piece of information as we begin to research immigrant populations. Accordingly, in the United States, the growing foreign-born community—particularly the diverse Hispanic groups—needs to be distinguished from the historic domestic populations such as Mexican-Americans and Puerto Ricans. This is particularly important in light of the powerful effect of socioeconomic position.

Class. Class or socioeconomic position is best defined by level of completed education. The lower the educa-

tional level, the lower the income and the less likely the individual is to have stable and continuous employment, which in turn influences access to health resources and benefits. Twelve years of education or fewer are highly correlated with fewer economic resources; more traditional knowledge, attitudes, and beliefs; fewer employment opportunities; and less likelihood to have health insurance and to use timely and appropriate preventive health care services.

An analysis of initiation of prenatal care by educational level shows that only about half of women with fewer than 12 years of education initiate care in the first trimester. This pattern is similar across ethnic and racial groups when education is controlled for (Brown, 1988). On the other hand, Tajali-Queiro (1989) found in her study of middle-income Hispanic women that 80 percent initiated care in the first trimester and continued to visit health care providers at a level similar to other groups in the population. Thus, there is a need to distinguish class groups within cultural groups. One approach is to construct a combined or multiple measure of number of years of school completed, health insurance status, and employment status to obtain an accurate descriptor of socioeconomic position which also represents an assessment of access to health resources.

Psychosocial factors. Psychosocial factors represent a set of contextual and quality-of-life indicators that are being recognized as critical in understanding health status, use of health services, and medical outcomes. My goal is not to discuss the intricacies of psychological measurement but rather to propose some areas of measurement that would enhance our understanding of the health of low-income and racial and ethnic populations. It has been well documented that low-income individuals experience higher levels of depression, more life events, and lower levels of social support, particularly instrumental support.

Simply put, life problems or chronic stress and the different types of social support available—family, friends, partner, and social welfare benefits—have a significant influence on the lives and health status of poor and racial/ethnic individuals in this society. These variables must be measured to better understand the context in which patterns of health care utilization and medical outcome occur. Lastly, lifestyle behaviors—which include use of alcohol, drugs, and cigarettes—constitute another critical dimension in understanding medical outcome.

Health Care Systems: Institutional Dimensions

Access to health care and quality of care are key institutional dimensions that clearly influence health outcome and medical effectiveness. The major variables to be measured are types of barriers, continuity of care, and quality of care. Factors such as organizational capacity—that is, the fact that patients have to wait 3-4 months for a prenatal care appointment, as has been the case in Los Angeles—cannot be ignored as a determinant in initiation of prenatal care or birth outcome. The lack of health insurance, which contributes to low use and delayed use of prenatal care, clearly influences medical outcome. The 1-minute garbled communication with a physician or lack of communication due to language barriers influences compliance with a medical regimen and medical effectiveness. Thus, one cannot assess medical outcome without examining the institutional context in which the medical interaction occurs.

Dimensions of a Conceptual Paradigm

Usually a paradigm or conceptual model serves to guide the work in a substantive area of inquiry. There appears to be an emerging knowledge base that enables health services researchers to formulate a multidimensional model for addressing the diverse health care problems in the United States. The impetus for this development is the diverse population changes that have occurred in the last 15 years. Thus, the emerging paradigm, I believe, points to the need for a consensus on some basic sociodemographic indicators to measure precisely race, ethnicity, nativity, and number of years in the United States. In addition, socioeconomic indicators—such as number of years of school completed, employment, and insurance status—may represent a combined proxy indicator for social class. A second set of factors needed to guide the model is related to the measurement of psychosocial indicators of chronic life problems; availability of family, friends, partner, and organizational supports; and lifestyle behaviors. The third set of factors is related to institutional and quality-of-care dimensions such as barriers to health care, language of provider, ability to buy prescriptions, and physical and financial accessibility of a facility—for instance, does the facility take Medi-Cal or is there a self-pay system?

These dimensions have repeatedly been set forth as important (see, for example, Aday, Fleming, and Andersen, 1984; Aiken and Mechanic, 1986; Andersen and Andersen, 1979; Elinson and Siegmann, 1979; Zambrana, 1988). What is innovative, however, is the scientific measurement of multiple factors to assure specific ethnic and class identifiers and the specific dimensions that need to be included in a study if we are to understand better the interrelationships among health status, health care utilization patterns, and medical outcome in special populations. These areas represent the dimensions to design cross-culturally sensitive studies and measure those variables that will contribute to our

knowledge base to improve medical practice and the health status of the U.S. population.

Methodological Issues in Using Medical Outcome Records

Existing data collection procedures and data systems present unique problems in using secondary data sources, such as medical records. This section describes our experiences over the last 2 years in abstracting the medical records of 1,234 low-income, racially and ethnically diverse, pregnant respondents in Los Angeles County.

The aim of our two-stage study was to examine ethnic differences in birth outcomes in primiparous low-income women in Los Angeles. The data collection approach consisted of face-to-face interviews, using a survey instrument, with 1,591 respondents after 20 weeks of pregnancy and subsequent retrieval and abstraction of the labor and delivery medical record data.

Identification of eligible survey respondents. The criteria for selection of our respondents were:

1. Age 17-35;
2. Parity, with no prior pregnancies greater than 16 weeks;
3. First birth (primiparous) with the intention to carry to full term;
4. Ethnicity defined for each of four groups: whites born and raised in the United States of white parents; blacks born and raised in the United with both parents of black heritage; Mexican-Americans born and raised in the United States or raised in the United States since the age of 10 with both parents of Mexican origin; and Mexican immigrants born and raised in Mexico of parents of Mexican origin and in the United States for 7 years or fewer; and
5. Educational level of 12 years of completed education or less and Medi-Cal and/or self-pay status served as identifiers of low socioeconomic status.

Survey. In total, the main survey study took 26 months; 45 sites were approached, and 22 sites were used for the survey study. Eligibility screeners, using prenatal clinic charts, filled out medical record abstraction forms for 5,336 potential respondents, out of which 1,592 were eligible and interviewed. The principal reason for ineligibility was noneligible ethnicity—that is, a significant number of Central Americans, mainly from El Salvador and Guatemala, were using public health facilities. Educational level disqualified 7 percent of the potential respondents. This was insightful to the extent that it reaffirmed that Medi-Cal status alone cannot be used as a socioeconomic indicator. These individuals were Medi-Cal recipients; yet their educational level ranged from college to masters' degrees (N = 262).

Another interesting finding was that although our overall refusal rate was low at 4.7 percent (N = 75), an analysis of interethnic refusal rates showed that 59 percent of the respondents who refused to participate were Mexican-American (N = 44), followed by black (17 percent) and Mexican immigrant (7 percent). We attribute the higher refusal rate of the Mexican-American respondents to the fact that the majority were English speaking and did not want to struggle with an interviewer who did not command the English language. Again when, interviewers were matched with Mexican-Americans, cooperation was high. Overall, there were 25 duplicates, a significant number being Mexican immigrant—that is, respondents who agreed to be interviewed twice without advising the interviewer. Because of the length of the interview (approximately 1 hour for Spanish), we speculate that these respondents enjoyed talking with the interviewers while waiting for their appointment or were afraid not to participate, since interviews were conducted in public clinics.

Furthermore, it was found that several individuals classified as black respondents (N = 14) were from Haiti and other Caribbean countries, and both Asian and Native American respondents (N = 8) were classified as Hispanic due to classification of Spanish-sounding last names as a proxy for ethnicity. These data provide additional evidence that patient medical records may contain significant errors that can compromise the ability to identify special populations. (For a description of the survey stage of the study, see Zambrana, 1991.)

Medical record abstraction procedures. The medical records abstraction form was designed to provide a systematic and simple codified tool to abstract maternal medical risk factors that influence both maternal and neonatal outcomes. The form was developed using Problem Oriented Patient Risk Assessment (POPRAS) forms 1-8 to identify relevant and available data. The format included references to already existing codes on the POPRAS in a sequential manner to assure simple and efficient abstraction of data. The form proved useful, but not all hospitals had medical forms that were consistent with the sequence. This became particularly problematic in abstracting delivery data, since significant information was written in notes.

Ample evidence suggests that medical records have serious limitations for evaluating medical effectiveness, such as missing data and illegibility, unless a clinic or program is specifically designated as a research center, such as the Pediatric Pulmonary Centers. This discussion addresses methodological caveats in the identifica-

tion and use of medical records for maternal and child health outcome data with low-income and racial/ethnic populations. Medical record abstraction presented a series of methodological issues during the course of the study which are simply divided into institutional dimensions and respondent dimensions.

Institutional dimensions. A discussion of hospital factors is related to both personnel issues and errors in data recording. The major institutional barrier was related to medical record department personnel being understaffed, overworked, and overwhelmed with little time to check carefully the identifiers to verify records. In all instances, medical record personnel were courteous and cooperative. The more understanding we showed, the more they tried to help us.

However, as a result of high staff turnover, including heads of departments, medical records were retrieved at a significantly slower pace than originally expected—that is, about 40-50 medical records per visit. Most departments requested that we submit a list of no longer than 25 or 30 names, and in most instances, only about 50 percent were retrieved. In general, we found that the number of charts pulled varied tremendously from site to site and from week to week. We often had to wait several weeks for a site to pull the charts requested. Often, there were extensive delays due to either other investigators (in house) making requests or major medical quality assurance audits being conducted. Thus, this phase of the project took about 18 months as opposed to the projected 6 months.

Initially, it was agreed that the abstraction would be conducted under a subcontract to the research firm that would train an abstractor. As discussed above, the actual access and retrieval of the records did not constitute a smooth process whereby a medical record abstractor could be told to work for certain hours. Second, availability of records was usually confirmed the day before, and more often than not, the schedule was changed. Thus, 2 months into this phase of the project, it was deemed necessary to hire the project director full time to coordinate this effort and two physicians (who were in a Master of Public Health graduate program) to decipher the records. Lastly, our costs were projected on each chart being abstracted in approximately 15-20 minutes. Our experiences quickly proved us wrong. It required from 20 minutes to 1 hour to abstract a medical record. The minimal time was for those records in one facility that had all the medical record forms typed.

The medical record itself was also replete with errors that had to be continuously checked, such as incorrect names in different parts of the record, differences in the utilization of last names within the Hispanic culture (for example, use of husband's last name and maiden name), and incorrect date of birth (DOB). These latter issues are related to respondent dimensions.

Respondent dimensions. One of the greatest problems we faced was that less than half of the 1,591 women interviewed actually had charts at the sites they had originally reported during their face-to-face interview. In many of the public prenatal care county facilities, what is referred to as "ob contractors" would talk with the pregnant patient at or about 37 weeks of pregnancy to discuss and confirm the delivery site. This system served to assure a more balanced distribution of deliveries in the Los Angeles County system—which maintains subcontracts with a large number of other private hospitals—and to enable the clinics to monitor and followup their prenatal patients. These subcontracts, however, may change at any time (they may be canceled, or new hospitals may be added). Furthermore, many of the respondents, especially English-speaking respondents, did not want to deliver at assigned County facilities.

We used four major indicators to retrieve medical records: name, DOB, expected date of delivery (EDC), and sometimes prenatal care clinic (since there appeared to be a referral system). Chart numbers were also recorded and maintained as a major identifier. It soon became clear, however, that although certain public clinics did have a system for specific county hospitals to use the clinic number as the medical record number, this procedure was not consistent across clinics. Fortunately, several months into the study, we began to request respondent home phone numbers and permission to verify their delivery site later in the pregnancy.

During the first 6 months of the second phase of this project, we were unable to understand why the medical records were not at the expected delivery site. We then initiated telephone calls to approximately 600 respondents. We spoke with respondents and friends and relatives of respondents and obtained some very insightful information, especially with regard to the Hispanic population.

During the telephone conversations, we requested confirmation of last names used at delivery site, DOB, and actual site of delivery. Frequently, the actual site of delivery was incorrect or we had submitted the record to another site at which the respondent reported to have delivered. The information that was incorrect, for the Hispanic respondents, was usually the last name, for several reasons; one, because the facility did not know that both maternal and paternal names were used and usually selected the first one (maternal name); two, the woman had used her partner's name at delivery; three, the respondent had left the baby's father and was using her own maiden name; or four, the respondent had a new partner and was using his name.

The second area of verification was DOB, which hospitals use especially to distinguish members of Hispanic groups who tend to have similar first and last names. It was discovered that the respondents did not know their exact DOB or had forgotten what DOB they had originally given. We tried to have respondents tell us the range of possible dates they had given and would then proceed to submit the retrieval request several times with different DOBs. Different ways of writing the DOB was also a point of confusion. For example, in Latin American countries, the day is first, followed by the month and the year (22-8-91). In both data entry and record retrieval, this initially caused confusion.

We also encountered a series of other dilemmas. In some cases, the women's names and DOBs were correct, but the records did not have labor and delivery dates—we found that it was not our respondent. Often, records were not available because of a second pregnancy or the records were in medical audit. At other times, a facility had a medical record on one of our respondents, but it was for a medical emergency visit and not a delivery record. Thus, it became clear that although the respondent used that particular site, she did not deliver there. After further investigation, we discovered that County facilities were so overcrowded that women who arrived in early labor were often transferred to other subcontractor facilities; the place of transfer was not always available. Lastly, EDC was sometimes inaccurate by 1-2 months. For those facilities where EDC was a major verification indicator, records were usually not found unless we obtained the correct EDC from respondent.

The identification of these problems led to the development of extensive medical record "search-and-find" procedures to match respondent and site of delivery. Once we established that the delivery records were not at the reported facility, the cases that we now refer to as "missing" were searched for at other hospitals in the county. We assumed that since the women were low income and had few choices for delivery site, they were most likely to deliver at one of the three major county delivery sites.[1] Therefore, the list of respondents was circulated to each and every major facility to retrieve records. A second approach involved a return to the sites of prenatal care to determine from their records where the women actually delivered. At one study site, for instance, when the interviewer was waiting for eligible women to come for their appointments, the interviewer confirmed delivery information for women we had interviewed there.

[1]In this study, only one-third of the total sample had Medi-Cal; however, black women were the most likely to have Medi-Cal (64.3 percent), followed by Mexican-Americans (47.2 percent) and Mexican immigrants (13.2 percent).

These approaches enhanced our matched medical record rate. A total of 1,591 women were interviewed; we matched records for 1,234 respondents. Overall, we matched 78 percent of our survey data with medical record data. Ethnic match rates were highest for Mexican-Americans and Mexican immigrants (79 percent in each group), African-Americans (76 percent), and whites (68 percent).

Finally, 26 hospital sites were used to retrieve medical records. Three large county facilities yielded 46.6 percent of the medical record data. Three private (subcontract) county facilities yielded 22.2 percent, 8.2 percent, and 6.0 percent. Thus, 6 hospital sites provided 83 percent of all medical record data; the other 20 hospitals yielded 17 percent of the data.

Overall, we expected a 20-percent experimental loss in the study population based on the estimated rate for low-income populations in Los Angeles. Indeed, some of the respondents were lost to the study because they had moved prior to delivery. In some cases, the respondents had returned to Mexico or had moved out of State. We were able to confirm approximately 25 cases in which women had moved prior to delivery; however, approximately 350 women no longer lived at their original residence and had left no forwarding address.

Overall, these data help to identify the issues related to the new concept of "hard-to-reach populations." These issues are related to the socioeconomic status of this group in society—in particular, instability of housing arrangements; respondents tend to live with partner, family, or friends. There also appears to be residential mobility within this extended kin-and-friend network. Telephones are frequently disconnected or changed due to this mobility, and respondents tend to have a persistent pattern of episodic and discontinuous care. Furthermore, there seems to be extensive confusion regarding the last name as well as changes in last name because of changes in partners and exact date of birth. These respondent and institutional dimensions have to be addressed to assure that medical effectiveness outcome data are valid and reliable, especially in linking respondent medical data over time.

A Recommended Set of Research Principles to Improve Data Quality

Any recommendation to improve data quality must be embedded in the larger institutional and national health care context. Under the current health care system, persistent patterns of low use of health care services, and delayed or episodic care in low-income, racial, and ethnic populations, make the measurement of medical outcome effectiveness a difficult and complex task. Preliminary steps are needed before any effective medical

outcomes research is begun. The following areas represent a set of research principles for consideration and implementation to improve data quality:

1. The health services and medical research communities need to arrive at a consensus on operational definitions of identifiers to distinguish poor and racially, ethnically, and culturally distinct populations in the United States.

2. There is a need for collaborative and multidisciplinary research development with community health care providers who know the needs and concerns of the study group.

3. Instrument and measurement development needs to be guided by a conceptual paradigm (consensus on dimensions to be measured) to assure the appropriateness of the variables to be measured for target groups under study.

4. Knowledge of the organizational context of health services, such as issues of accessibility and availability of services, resurfaces as critical in a system such as the one in Los Angeles where the average wait for an initial prenatal care appointment is 12-16 weeks. The importance of measuring institutional dimensions or knowing the local context of services before jumping to conclusions is critical.

Thus, the issue remains of whose responsibility it is to abide by good scientific principles. What are the sanctions? What are the rewards? The concept of cognitive studies currently being proposed by NCHS perhaps captures the essence of what needs to be done. The issue now is how to disseminate this new technology to investigators and how to hold them responsible. Thus, a reasonable future solution for generating knowledge on poor and culturally and racially distinct populations by region is to develop a guiding conceptual paradigm that incorporates comparable and standardized instruments using relevant and appropriate standard operational definitions of ethnic, racial, and class identifiers across data collection systems.

Management Data Systems: How Do We Initiate and Implement These Solutions?

The development of viable and useful data management systems to collect compatible and supplementary data at the national, State, and local levels is an effort long past due and certainly one that can be exceptionally useful—for example, the Tumor Cancer Registry. In my view, the initiative for this effort needs to be mandated by the Federal Government through its primary vital and health statistics agency, the NCHS. This agency has the opportunity to take a leadership role, through its work with other Federal agencies and academic institutions, to forge and implement a consensus on indicators for special populations as a national priority. At present, NCHS conducts four national population-based surveys, five record-based surveys, and four national vital statistic system surveys throughout the United States.

A second avenue is the National Vital Statistics System which provides vital statistics through State-operated registration systems. The procedures for uniform registration practices and use of records for national statistics are established through cooperative arrangements between States and NCHS. NCHS incurs the costs of the States in providing vital records for national use (DHHS, 1988); yet the availability and uniformity of State data for local and regional planning and research purposes have, for the most part, been fraught with difficulties.

Our own experiences in Los Angeles most likely mirror the chaos in other large regional areas. Currently, health care providers in the Los Angeles health care system have limited data on the populations they serve. With respect to availability of data, we were informed by several State and local administrators and maternal and child health researchers that although data from the POPRAS form is input into a computer system by many county institutions, these data are not available to the public. Nonetheless, an enormous amount of time was initially spent attempting to obtain data that were discovered to be unavailable. In the end, actual site visits, discussions with medical chiefs and clinical nursing administrators, and feasibility studies at the study sites were conducted to determine the social and demographic characteristics of the populations seen at each site.

In addition, there is no system for the actual management of pregnancy through labor and delivery. For example, when a woman delivers at a hospital, she usually has no medical record to review if she has been using a county clinic or a private physician, unless she has been given a copy of her medical record and is carrying it with her. This clearly curtails any reliable data on outcome.

This finding is not new; the problem has been clear for several decades and was highlighted in the 1970s. Currently, the Robert Wood Johnson Foundation has a "new initiative" that "intends to help States strengthen their health statistics systems to support policy making, as well as to aid in program development and management." The California Chief of the Health Demographics Section states:

> Across California, many agree that the current health data infrastructure in the State is inadequate to address today's critical health policy issues. This inadequacy is seriously undermining our State's, as well as our Nation's, abilities to make policy deci-

sions. Since the 1970s, there has been an increased need for timely and accurate health data. During this same period, there has been a decreased capacity to meet these needs.[2]

The collection of data by States is crucial for a better understanding of patterns of morbidity and mortality by geographic location, in context, particularly the allocation of health resources. For example, the percentage of individuals in the State covered by Medi-Cal and WIC provides a context in which to interpret medical outcomes data.

In addition to the State data system, two other viable mechanisms for the development of local data management systems can be linked to and supplement both State and national data systems: the first is a network of practice-based research sites to enhance our data bases by select geographic regions (for example, inner city populations or seasonal migrant farm workers in rural areas or poor whites in Appalachia [see, for example, Christoffel, Binns, Stockman, and others, 1988]). During our study, we were informed of a network of black physicians who serve the local residents of surrounding black communities. These types of networks need to be further explored as potential sources of valuable information. In our experiences, these physicians have proven to be exceptionally cooperative, committed, and interested in research to better understand their patient populations.

Another mechanism is the community health center, which is a prime vehicle for the development of family-based data systems. I was surprised recently when I was asked to serve on the Comprehensive Perinatal Care Program (CPCP), implemented in 1988 by the Bureau of Health Care Delivery and Assistance to support services designed to improve pregnancy outcomes and health status of poor and medically underserved women and infants. The CPCP was, at the time of the writing of this paper, devising data strategy objectives for program/planning and management and evaluation/assessment of outcomes. Its study of the data capacity of community and migrant health centers revealed that either the systems were manual and costly or highly technical and unfriendly. In sum, the data capacity system is inadequate and needs revamping. CPCP also concluded that supplemental sources of data such as those maintained by NCHS are needed to supplement center data (CPCP, 1991). This report revisited the same discussions and issues that were raised in 1973 in the area of the importance of developing adequate data management systems in community health centers.

The 1970s represented the heyday of the management information system as a powerful tool for the collection of data that would be representative of local needs and assist in planning and policy. I am somewhat struck by the need of the Federal Government to commission a study on data systems in these centers, since, if my memory serves me, when I was developing the grant application for a neighborhood health center in East Harlem, there were no financial resources available for a data collection system to evaluate the centers.[3] As was stated 20 years ago, community and migrant health centers represent crucial organizational sites for the collection of health status and medical outcomes data. They should indeed be funded to collect these data and work closely with NCHS to assure comparability of data across sites. It seems that there are no new suggestions, but we do need to put into practice what we already know and what makes good sense.

In the end, our medical effectiveness research can only measure how well the health care system has responded to the needs of its citizens. The "lens" or world view of researchers will continue to capture only a small snapshot of the picture unless they view research as a scientific tool to open up new worlds of understanding and not simply replicate the old stereotypic notions of the poor as miscreants.

It appears that the national capability exists to revise major data collection systems that can better serve the needs of special populations in the United States. What does not exist is the sense of purpose. An indepth review of the types of data collection systems we have in place is astounding; yet the data have failed to help us address the problems that have plagued our health system for decades. Sultz (1991) discusses this issue with clarity in reflecting on our health care system:

> It allows us to collect masses of data about the details of health care without addressing authentic issues of comprehensiveness, continuity, and effectiveness. As our medical care technology becomes more sophisticated, the more primitive and indifferent the system seems to be in meeting the actual health needs of our society. We are moving with agonizing slowness toward the kinds of outcomes research that can lead to authentic understanding of the problems of health care and effective remedial action. (p. 419)

[2]Letter of May 21, 1991 from Gwendolyn Doebbert, Chief, Health Demographics Section, Department of Health Services, State of California, to David Hayes-Bautista, Ph.D., Director, Chicano Studies Research Center, University of California, Los Angeles.

[3]For a discussion of neighborhood health centers, see: Davis, K. and Schoen, C. (1978). *Health and the war on poverty.* Washington: The Brookings Institution.

References

Aday, L. (1989). *Designing and conducting health surveys: A comprehensive guide.* San Francisco: Jossey Bass.

Aday, L., Fleming, G.V., and Andersen, R. (1984). *Access to medical care in the U.S.: Who has it, who doesn't.* Chicago: Pluribus Press, Inc., and the Center for Health Administration Studies, University of Chicago.

Aiken, L.H. and Mechanic, D. (1986). *Applications of social science to clinical medicine and health policy.* New Brunswick, NJ: Rutgers University Press.

American Medical Association. 1991. Hispanic health in the United States. *Journal of the American Medical Association,* 265(2), 248-252.

Andersen, R. and Andersen, O.W. (1979). Trends in use of health services. In Freeman, H., Lenen, S., and Reeder, L.G. (Eds.), *Handbook of medical sociology* (pp. 371-391). Englewood Cliffs, NJ: Prentice Hall.

Becerra, J.E, Hogue, J.R., Atrash, H.K., and Perez, N. (1991). Infant mortality among Hispanics: A portrait of heterotgeneity. *Journal of the American Medical Association,* 265(2), 217-221.

Boone, M. (1989). *Capital crime: Black infant mortality in America.* Newbury Park, CA: Sage.

Byrd, W.M. (1990). Race, biology, and health care: Reassessing the relationship. *Journal of Health Care for the Poor and Underserved,* 1(3), 278-296.

Chesney, A.P., Chavira, J.A., Hall, R.P., and Gary, H.E., Jr. (1982). Barriers to medical care of Mexican-Americans: The role of social class, acculturation, and social isolation. *Medical Care,* 20(9), 883-891.

Christoffel, K.K., Binns, H.J., Stockman, J.A., III, and others. (1988). Practice-based research: Opportunities and obstacles. *Pediatrics,* 82(3, Part 2), 399-406.

Cousins, L. (1987). Pregnancy complications among diabetic women: Review 1965-1985. *Obstetrical and Gynecological Survey,* 42(3), 140-149.

CPCP Perinatal Data Workgroup. (1991). *Background briefing materials for Washington, DC workgroup meeting, May 22-23, 1991.* Rockville, MD: Health Resources and Services Administration.

Cramer, J.C. (1987). Social factors and infant mortality: Identifying high risk groups and proximate causes. *Demography,* 24(3), 299-322.

Cuellar, I., Harris, L.C., and Jasso, R. (1980). An acculturation scale for Mexican American normal and clinical populations. *Hispanic Journal of Behavioral Sciences,* 2(3), 199-217.

Department of Health and Human Services. (1986). *Report of the Secretary's task force on black and minority health. Volume 4: Infant mortality and low birth weight.* Washington: Author.

Department of Health and Human Services (DHHS). (1988). *Organization and activities.* Hyattsville, MD: National Center for Health Statistics.

Deyo, R.A., Diehl, A.K., Hazuda, H., and Stern, M.P. (1985). A simple language-based acculturation scale for Mexican-Americans: Validation and application to health care research. *American Journal of Public Health,* 75(1), 51-56.

Doherty, W.J. and Campell, T.L. (1988). *Families and health.* Newbury Park, CA: Sage.

Eberstein, I.W. (1989). Demographic research on infant mortality. *Sociological Forum,* 4(3), 409-422.

Eberstein, I.W., Nam, C.B., and Hummer, R.A. (1990). Infant mortality by cause of death: Main and interaction effects. *Demography,* 27(3), 413-430.

Elinson, J. and Siegmann, A.E. (Eds.). (1979). *Sociomedical health indicators.* New York: Baywood.

Ginzberg, E. (1991). Access to health care for Hispanics. *Journal of the American Medical Association,* 265(2), 238-241.

Grady, K.E. and Wallston, B.S. (1988). *Research in health care settings.* Newbury Park, CA: Sage.

Hayes-Bautista, D.E. (1990a). *Latino health indicators and the underclass model: From paradox to new policy models.* Los Angeles: University of California, Chicano Studies Research Center.

Hayes-Bautista, D.E. (1990b). *Latino family policy: An explanatory framework from California.* Los Angeles: University of California, Chicano Studies Research Center.

Hayes-Bautista, D.E., Hurtado, A., and Valdez, R.O. (1990). *Redefining California: Latino social engagement in multicultural society.* Los Angeles: University of California, Chicano Studies Research Center.

Health Care Financing and Organization. (1991). *News and progress* (p. 8-9). Washington: Alpha Center.

Health status among Hispanics: Major themes and new priorities [Editorial]. (1991). *Journal of the American Medical Association,* 265(2), 255-257.

Hispanic health: Time for data, time for action [Editorial]. *Journal of the American Medical Association,* 265(2), 253-255.

Institute of Medicine. (1988). *Prenatal care: Reaching mothers, reaching infants.* Washington: National Academy Press.

Johnson, D. (1991). Who should be responsible for social responsibility in research? *Psychological Science*, 2(3), 59-61.

Marin, G. and Marin, B. (1991). *Research with Hispanic populations*. Newbury Park, CA: Sage.

Mendoza, F.S., Ventura, S.J., Valdez, R.B., and others. (1991). Selected measures of health status for Mexican-American, mainland Puerto Rican, and Cuban-American children. *Journal of the American Medical Association*, 265(2), 227-232.

Neresian, W.S. (1988). Infant mortality in socially vulnerable populations. *Annual Review of Public Health*, 9, 361-377.

Olmedo, E.L. and Padilla, A.M. (1978). Empirical and construct validation of a measure of acculturation for Mexican-Americans. *Journal of Social Psychology*, 105, 179-187.

Rogers, R.C. (1989). Ethnic and birth weight differences in cause-specific infant mortality. *Demography*, 26(2), 335-343.

Schorr, L.B. (1989). *Within our reach: Breaking the cycle of disadvantage*. New York: Doubleday.

Sirken, M.G., Bercini, D.H., and Jobe, J.B. (1990). National Center for Health Statistics Data Line. *Public Health Reports*, 105(5), 538-539.

Stroup, C.A. and Trevino, F.M. (1991). Reproductive characteristics of Mexican-American, mainland Puerto Rican, and Cuban-American women. *Journal of the American Medical Association*, 265(2), 222-226.

Sultz, H. (1991). Health policy: If you don't know where you're going, any road will take you. *American Journal of Public Health*, 81(4), 418-420.

Tajali-Queiro, I. (1989). Hispanic women's perceptions and use of prenatal health care services. *Affilia: Women and Social Work*, 4(2), 60-72.

Trevino, F.M. (Ed.). (1990). Hispanic Health and Nutrition Examination Survey, 1982-84. *American Journal of Public Health*, 80(Suppl).

Trevino F.M., Moyer, M.E., Valdez, R.B., and Stroup-Benham, C.A. (1991). Health insurance coverage and utilization of health services by Mexican Americans, mainland Puerto Ricans, and Cuban Americans. *Journal of the American Medical Association*, 265(2), 233-237.

Warner, P.C. (1991). Health issues at the U.S.-Mexican border. *Journal of the American Medical Association*, 265(2), 242-247.

Zambrana, R.E. (1988). A research agenda on issues affecting poor and minority women: A model for understanding their health needs. *Women and Health*, Winter, 137-160.

Zambrana, R.E. (1991). Cross-cultural methodological strategies in the study of low-income racial ethnic populations. In Grady, M.L. (Ed.), *Primary care research: Theory and methods*. Rockville, MD: Agency for Health Care Policy and Research.

Acknowledgments: The experiences and data cited in this paper are drawn from a project (HS05518) funded by the Agency for Health Care Policy and Research. The author acknowledges the effort and contributions of Drs. Susan Scrimshaw and Christine Dunkel-Schetter, Co-Principal Investigators; Nelly Valencia, Project Director; and Frantz Wilson, Project Officer, Agency for Health Care Policy and Research.

Dilemmas in Measuring Medical Outcome in Low-Income and Racial/Ethnic Populations

Jose J. Escarce, M.D., A.M.

Dr. Zambrana's paper (in this volume) indicates a desire to improve research and data regarding low income and minority populations in the United States. Her main premise is that our understanding of health care utilization and health outcomes in these populations has been hampered by a failure to take into account the social environment and psychosocial and family contexts in which disease develops and use of health care services takes place. As a solution to this problem, the paper proposes a conceptual paradigm, or model, to guide research on racial and ethnic differences in health services utilization and health outcomes.

The other theme of the paper pertains to the difficulty of collecting data of high quality on low income and minority populations. The author relates her own recent experiences during a study of low income and minority pregnant women in Los Angeles, experiences that vividly illustrate the lack of applicability of the usual assumptions regarding the behavior of patients and the performance of health care institutions. The paper concludes with the following recommendations for improving the quality of data on low income and minority populations: develop a consensus on operational definitions of identifiers to distinguish racially and ethnically distinct populations in the United States, develop measurement instruments and data collection procedures that are appropriate to the groups under study, and take careful account of accessibility and other institutional dimensions of the health care system in collecting and analyzing the data.

Several of the issues raised in this paper merit close attention. As Dr. Zambrana points out, historically there has been avoidance of certain minority populations in health services research. This observation especially pertains to Hispanic groups. Dr. Zambrana also emphasizes the importance of examining characteristics of the health care delivery system, such as accessibility and quality, in studies of racial and ethnic differences in utilization and outcomes. These factors, which sometimes are neglected by researchers, may have considerable explanatory power. Finally, as Dr. Zambrana suggests, instruments and data collection procedures that are appropriate to the groups under study are important to obtain valid and reliable data for minority populations.

The conceptual model presented in the paper is actually a list of variables to be measured and used in studies of health services utilization, health status, and health outcomes. The model comprises three different types of variables. The first type of variable consists of sociodemographic variables, which include race; indicators of ethnicity and level of acculturation; and education, health insurance status, and employment status as indicators of socioeconomic position. The second type of variable consists of psychosocial variables, which include measures of chronic life problems and the availability of family, friends, and institutional supports. The third type of variable consists of variables measuring institutional dimensions of the health care system and, in particular, access to care and quality of care.

Two general comments are pertinent regarding this model. First, from an empirical perspective—that is, viewed as a list of variables to be measured and used in empirical research—the model is not new. As the author acknowledges, these same variables, with minor modifications, have been used in research on health services utilization and health outcomes for a number of years. It is also worth noting that, for certain types of studies, additional variables would be necessary. For example, comprehensive measures of health status are important in studies of the use of health services.

Second, from a theoretical perspective, the exposition of the model in the paper suffers from the lack of a consistent theoretical framework that would allow a clear explication of the mechanisms by which the variables in the model affect utilization and outcomes. Developing

Dr. Escarce is Assistant Professor, School of Medicine, and Senior Fellow, Leonard Davis Institute of Health Economics, University of Pennsylvania.

such a framework is not difficult. Several models in the literature have a well-specified theoretical basis and share variables with the model proposed by Dr. Zambrana.

For example, Dr. Zambrana's model bears a close relationship to the behavioral model of health services utilization developed by Andersen and colleagues (Aday and Andersen, 1975; Andersen, 1968) and used by numerous health services researchers. Briefly, the behavioral model views the use of health services as a function of the predisposing, enabling, and need characteristics of individuals. Predisposing characteristics are important because different individuals have different propensities to use health services. Predisposing characteristics are subdivided into three dimensions: demographics, such as age and sex; social structure, such as education and ethnicity; and health beliefs. Enabling characteristics are important because, to use health services, individuals must have the means to obtain them. Enabling characteristics are subdivided into two dimensions: individual or familial resources, such as income and health insurance status, and community resources, such as physician and hospital-bed supply. Finally, need characteristics are important because individuals generally use health services only when they have or perceive some illness. In the behavioral model, need is the most immediate cause of health services use.

An alternative theoretical framework is provided by the standard economic model of the demand for health services, which also shares many variables with Dr. Zambrana's model. In the economic model, however, these variables are viewed as measures of patient preferences; benefit from health care; socioeconomic status; and the money and time prices of health services, complements, and substitutes. The presentation of the model in the paper could be improved by highlighting its relationship to these or other models in the literature.

Irrespective of the theoretical framework chosen, however, the use of race and ethnicity as explanatory variables in studies of health care utilization and health outcomes requires comment. Using race and ethnicity as variables for taxonomic purposes is well accepted in social science research. These variables also have a long tradition of use in health services research, especially in studies of health beliefs, health knowledge and awareness of symptoms, help-seeking behaviors, use of health services, and access to health care. Race and ethnicity appear in the behavioral model of health services utilization as predisposing characteristics. In the behavioral model, these variables are mainly intended to capture an individual's location in the social structure and the behavioral patterns or lifestyles to which people in such positions become socialized (Wolinsky, Aguirre, Fann, and others, 1989). Similarly, race and ethnicity are often included in economic models of the demand for health services. In economic models, these variables usually are regarded as components of socioeconomic status.

In view of this long tradition, the notion that using race and ethnicity as explanatory variables needs discussion may seem surprising. Yet this practice raises conceptual and methodological issues which are often ignored by investigators but have major implications for data collection and analysis and for the interpretation of research findings.

As reviewed by Wilkinson and King (1987), the first difficulty with using race and ethnicity as explanatory variables arises because racial and ethnic classifications encompass a broad range of biologic and nonbiologic phenomena. At the most basic level, race can be conceptualized as a biologic or genetic category that has phenotypic correlates. Under this conceptualization, the race variable captures biologic factors such as gene pools and genetic susceptibility to particular diseases. However, race also captures many nonbiologic factors. For instance, race can be conceptualized as a social construct. Under this conceptualization, the race variable captures factors related to societal stratification, including the position of different groups in the social structure, segregation, and discrimination. The use of race as a variable in the behavioral model of health services utilization most closely fits this conceptualization. Race also can be conceptualized as a cultural category. Under this conceptualization, the race variable captures behavioral and cultural factors such as beliefs, values, and shared standards for behavior. Finally, race can be conceptualized as a measure of socioeconomic status. Under this conceptualization, the race variable captures factors related to economic resources, such as income and financial assets. This conceptualization of race is often adopted in economic models of the demand for health services.

Similarly, although in strict usage the concept of ethnicity refers to the shared social norms, cultural beliefs, and standards for behavior that result from common origins (Harwood, 1981), in practice ethnic classifications capture biologic, social structural, and socioeconomic factors as well. Physical traits and a genetic predisposition to certain diseases characterize many ethnic groups. Ethnicity, socioeconomic status, and social class also are closely interrelated in the stratified society of the United States.

These observations are relevant for research on racial and ethnic differences in health services utilization and health outcomes. To illustrate, consider the following simple model of utilization and outcome in an episode of illness. The episode begins with the onset of a disease,

regarded as an objective derangement in anatomy or physiology. The disease causes symptoms which, if perceived by the individual as illness, may lead the individual to seek medical care. The type and intensity of care, that is, the type and number of diagnostic and therapeutic services, may vary among individuals who have the same disease and symptoms. Likewise, health outcome may vary.

Race and ethnicity influence each phase of this process in varied and complex ways. The occurrence and severity of disease obviously may be affected by biologic or genetic factors; however, disease occurrence and severity also may be influenced by shared risk factors that are largely due to social class or socioeconomic status and by culturally determined pathogenic patterns of behavior. The perception of symptoms as illness and the decision to seek medical care may be affected by cultural beliefs and values, socioeconomic status, and the accessibility of treatment. Accessibility in turn is related to the organization and functioning of the health care system, which is profoundly influenced by social stratification and residential segregation. The type and intensity of care may be affected by socioeconomic status, cultural beliefs, and the organization of the health care system. Because providers partially control this phase of care, type and intensity of care also may be affected by inappropriate consideration of race or ethnicity in clinical decisionmaking, including discrimination. Finally, health outcome is a function of disease severity, type and intensity of care, and additional biologic factors that affect the response to treatment.

Researchers who perform studies of racial or ethnic differences in health care utilization or health outcomes generally attempt to control for other variables that affect utilization and outcomes using appropriate statistical techniques, but adequate control for all other potentially important variables is not possible. Some variables may not be recognized as influential; others may not be available to the researchers; and still others, such as discrimination, may be difficult to measure. Consequently, the "independent" effect of race or ethnicity observed in these studies is likely to be confounded with multiple and diverse biological and nonbiological factors. Moreover, the proportion of the effect of race or ethnicity attributable to each of these factors is undeterminable.

Failure to appreciate the breadth and complexity of the phenomena associated with race and ethnicity may result in simplistic and inappropriate interpretations of research findings. For instance, researchers may conclude that certain types of intervening factors, such as social or cultural factors, do not contribute to racial and ethnic differences in utilization or outcomes if including crude measures of these factors does not alter the estimated effect of race or ethnicity in a particular study. Alternatively, researchers may attribute the "independent" effects of race and ethnicity that remain after controlling for a small number of available variables to one or another set of intervening factors, often without clear empirical or theoretical justification.

The second difficulty with using race and ethnicity as explanatory variables arises because, in the usual approach, each racial or ethnic population is treated as a homogeneous group with respect to the effect on utilization or outcomes. Social scientists have identified several sources of intraracial and intraethnic variation in health beliefs and behaviors, including region of origin within the United States or country of birth, degree of acculturation to the norms of the dominant culture, and class differences (Harwood, 1981). For example, most research indicates that ethnic differences in health behaviors diminish among members of the upper socioeconomic classes. Intragroup variations in health care utilization and health outcomes highlight the essential role of intervening social, cultural, and economic factors in explaining the observed effects of race and ethnicity. Capturing these variations requires allowing potentially complex interactions among variables. Conversely, ignoring intragroup variations may obscure the importance of intervening factors and lead to erroneous conclusions regarding the causes of racial and ethnic differences. Neglect of intragroup variations also may exacerbate the tendency to stereotype members of racial and ethnic groups.

To conclude, as suggested by Dr. Zambrana, research on health care utilization and health outcomes in low income and minority populations in the United States is needed for several reasons. Descriptive studies can help to identify groups who are poorly served by our system of health care delivery and to document the magnitude of racial and ethnic disparities in utilization and outcomes. Analytic studies using multivariate statistical techniques can enhance our understanding of the factors that contribute to racial and ethnic differences in utilization and outcomes and may suggest corrective policies. Causal inferences, however, require cautious and careful interpretation of research findings. Recognition of the influence of race and ethnicity on every phase of the process of obtaining health care and of intraracial and intraethnic variations in health behaviors is essential to avoid simplistic and misleading conclusions. Comparative studies may fail to fully elucidate the relative importance of all the intervening factors that lead to racial and ethnic differences in utilization and outcomes.

Dr. Zambrana's recommendations for improving the quality of data on low income and minority populations

also are well considered, and some of these are already being addressed at the national level. Specifically, the Subcommittee on Health Statistics for Minority and Other Special Populations of the National Committee on Vital and Health Statistics is developing guidelines for identifying and coding race and ethnicity on all data that are collected as part of the national data collection system (Risa Lavizzo-Mourey, M.D., personal communication, August 1991). The Subcommittee is also examining whether national surveys should be modified or expanded when they are administered in certain geographic areas or situations to make the surveys more appropriate for minority populations.

Researchers who collect their own primary data also must make certain that their instruments and data collection procedures are appropriate to the groups under study. Institutional dimensions of the health care delivery system should receive special attention in data collection and analysis. Moreover, data on minority populations should include sufficiently detailed indicators of social class and socioeconomic status to allow examination of intragroup variations in health behaviors and outcomes. To avoid ad hoc explanations for observed phenomena, however, each step of instrument development and data collection and analysis should be guided by a well-specified and consistent theoretical framework.

Finally, Dr. Zambrana's suggestion to improve data management systems in community health centers and other local institutions deserves emphasis. In particular, improved data could help these institutions to identify and address the needs of their patients. Mechanisms to assure uniformity and comparability of data across sites could make local data useful adjuncts to data collected at the State and national levels, but achieving uniform data collection locally is likely to be a difficult endeavor.

References

Aday, L.A. and Andersen, R.M. (1975). *Development of indices of access to medical care*. Ann Arbor, MI: Health Administration Press.

Andersen, R.M. (1968). *A behavioral model of families' use of health services*. Chicago: Center for Health Administration Studies.

Harwood, A. (1981). Introduction. In Harwood, A. (Ed.), *Ethnicity and medical care*. Cambridge, MA: Harvard University Press.

Wilkinson, D.Y. and King, G. (1987). Conceptual and methodological issues in the use of race as a variable: Policy implications. *Milbank Quarterly*, 65 (Suppl. 1), 56-71.

Wolinsky, F.D., Aguirre, B.E., Fann, L. J., and others. (1989). Ethnic differences in the demand for physician and hospital utilization among older adults in major American cities: Conspicuous evidence of considerable inequalities. *Milbank Quarterly*, 67, 412-449.

Record Linkage: An Overview

Leslie L. Roos, Ph.D.
Andre Wajda, M.Sc.
J. Patrick Nicol, B.Comm.
Jan Roberts, M.D.

Introduction

Record linkage is the bringing together of information from two independent source records that are believed to relate to the same individual or family (Acheson, 1967). Such records are said to be "linked" and may be treated as a single record for one individual or family. With successive linkages, the information may take on the characteristics of a collection of personal or family histories.

The term "medical record linkage" was first used in 1946 by H.L. Dunn, Chief of the U.S. National Bureau of Vital Statistics (Acheson, 1968). As described by Acheson (1987), Dunn introduced the new term to a group of Canadian vital statisticians in this way:

> Each person in the world creates a book of life. This book starts with birth and ends with death. Its pages are made up of the records of the principal events in life. Record linkage is the name given to the process of assembling the pages of this book into a volume.

Dunn envisaged the assembly of health records into personal files over a lifetime, with linkage implying records brought together by means of common identification data (for example, name and date of birth) which ensure that records are assigned to the correct file (Acheson, 1967). By 1957, Canadian geneticist H.B. Newcombe had recognized the full implications of extending the principle to arrangement of personal files in family groups (Newcombe, 1957); in 1959, British researchers applied the computer to the problems of sorting and matching personal records with discrepant identification data (Acheson, 1967). Commenting on the first use of computerized medical records by a government medical officer in Britain, Godber (1968) noted, "It is not coincidental that he is about the only Medical Officer of Health in the country...who manages to get about 90 percent of the children immunized."

Key technical issues were identified fairly early. Smith (1984) summarized three major difficulties involved in the technical side of record linkage:

1. Using personal identifiers to discriminate "between the person to whom the record refers and all other persons in the population";
2. Deciding whether discrepancies in identifiers are due to mistakes in reporting for a single individual or to the presence of additional individuals; and
3. Processing the large volume of data necessary for record linkage while using a reasonable amount of computer time.

Canadian work on record linkage began in 1957 and built on good vital statistics records which were computer readable (Acheson, 1967). Early studies by Newcombe (1968) linked British Columbia health and vital records, showing that various source records could be linked and integrated in health care administration, in medical and demographic studies, and in the provision of health statistics.

More recently, interest in record linkage has grown markedly. Record linkage techniques can help identify the same patient for matching diverse files (hospital discharge abstracts, insurance claims, registries, vital statistics data) that contain similar identifiers (Shannon, Jamieson, Walsh, and others, 1989). Such techniques are increasingly important for expanding and correcting data bases used both in studying various diseases

Dr. Roos is Professor, Community Health Science, Faculty of Medicine, The University of Manitoba and Manitoba's Centre for Health Policy and Evaluation. Mr. Wajda and Mr. Nicol are Systems Consultants at the Manitoba Centre for Health Policy and Evaluation. Dr. Roberts is Medical Officer of Health, Manitoba Health and Ph.D. candidate, Department of Community Health Sciences, The University of Manitoba.

(Buehler, Berkelman, and Curran, 1989; Young, Roos, and Hammarstrand, 1991) and in assessing a wide range of treatments and procedures (Cohen and Duncan, 1988a; Roos, L.L., Sharp, and Wajda, 1989).

Major Uses of Linkage

Longitudinal studies of mortality. Across North America, record linkage for longitudinal studies of mortality has been carried out centrally—at Statistics Canada for Canadian mortality data and at the National Center for Health Statistics for American mortality data. This brief review concentrates on Statistics Canada because of our perception of the organization's leadership in linkage activities. Statistics Canada pioneered in the development and use of probabilistic record linkage, especially in the study of mortality after a particular occupational exposure. Howe (1989) outlined the typical design of such research:

> ...if one is following up a cohort of individuals with some specific exposure, ascertainment of mortality, for example, can be achieved by comparing the records for the cohort with mortality records for the geographic area in which the cohort resides.

As described by Fair (1990), the Statistics Canada group has focused on occupational and environmental health problems; of particular interest were "diseases that are delayed in their expressions, often by as much as two or three decades from the environmental insult to health that caused or contributed to their development." Combining record linkage techniques with a national data base has facilitated many studies of workplace hazards and cancer mortality based on long-term followup. Long-term followup has been carried out with a number of different occupational groups (asbestos workers, nickel miners, farmers, workers at Atomic Energy Canada, and so forth). Newcombe, Fair, and their collaborators use several types of records (Fair, 1990):

1. Starting-point records that "identify an individual as belonging to an exposed or at-risk population (or to a control population with which the other is to be compared)," and

2. End-point records that "identify an end effect, such as cancer and/or death."

Linkage techniques permit the searching of starting-point and end-point files to determine which individuals appear on both files. A third type of file may provide additional information (such as better or more identifiers) to aid in carrying out the record linkage.

The research group at Statistics Canada has advanced record linkage in a number of other ways. For example, the development of a standardized data collection package to aid in designing studies for later linkage with the Canadian Mortality Data Base promises to be useful in studies built around primary data collection (Carpenter and Fair, 1989). Recent studies have also focused on improving the information that names can provide in carrying out record linkage (Fair and Lalonde, 1991).

A comprehensive multifile data base. Clinical data collection and surveys are often episodic, depending on such factors as funding and investigator interest. Alternatively, administrative data bases are generated continuously and are characteristically available to researchers a few months after the end of the fiscal year. Their comprehensiveness suggests the possibility of large, longitudinal data bases encompassing several types of information.

The Oxford Record Linkage Study outlined an important goal of comprehensive population data bases, which is to have:

> ... a system of linked health records which brings together selected data of biological interest for a whole population commencing with conception and ending in death, in a series of personal cumulative files, the files being organized so that they can be assembled in family groups. The term record linkage may apply specifically to the techniques of assembling the files in spite of errors and omissions in the identifying particulars, or may be used in a more general sense to apply to the organization involved. (Acheson, 1967)

In Manitoba, the population health data base records all patient contacts with physicians, hospitals, and nursing homes. It was derived from information contained in the population registry and from health insurance claims routinely filed by physicians and health care facilities with the Manitoba Health Services Commission (MHSC) (Roos, L.L., Nicol, and Wajda, 1985). Information from the Manitoba Immunization Monitoring System, the Manitoba Cancer Treatment and Research Foundation, and Vital Statistics was incorporated into the data base. Seven files provide the bulk of the information contained in the Manitoba health data base (Roos, L.L., Nicol, Johnson, and Roos, N.P., 1979). Patient identification information is regularly included in the first six files; individuals within a family can be uniquely specified. The seventh file contains mortality information. Names and addresses are never available.

1. *Registration file.* This file contains data (back to 1970) on the insured population. The registry includes information on dates of coverage, place of residence, and date of death. Information on family structure has been used to infer marital status.

2. *Hospital file.* This file contains all inpatient and day surgery admissions (back to 1970) to acute and chronic care hospitals and includes information on: dates of admission and separation (that is, discharge, transfer, or death); limited information on services rendered (up to 16 diagnoses); all surgical procedures; identification numbers of the attending physician, surgeons, consultants, and anesthetists; and hospital identification number.

3. *Medical claims file.* This file is generated by the claims filed by physicians for payment and includes (back to 1970): date service was rendered; physician identification number; patient's residence; and tariff, determined by whether the contact is an office visit, consultation, outpatient contact, laboratory test, x-ray, or other service; and diagnosis.

4. *Personal care home file.* This file includes records (back to 1973) of dates of assessment, admission and separation of individuals in personal care homes, levels of care, and comprehensive data on drugs prescribed.

5. *Manitoba Immunization Monitoring System* (MIMS). Children born on or after January 1, 1980, are automatically entered into MIMS on receiving their first immunization. The MIMS file contains date of birth, sex, date of termination (death or migration), and attached immunization records (including immunization tariff code, service date, and restrictions).

6. *Cancer registry.* This file includes information (back to 1970) on date of cancer diagnosis and type of cancer.

7. *Mortality file.* This file includes information on date and cause of death.

Each of these files is maintained separately by its parent organization, with no routine record integration. Since patient and physician identification has been standardized across all files, however, files can be linked so that the data can be viewed from several different perspectives. Histories can be developed of individuals' contacts with physicians or hospitals over time, of all care rendered by physicians over time, or of hospital or personal care home admissions over time (Roos, L.L., Roos, N.P., Cageorge, and Nicol, 1982).

The Manitoba data base thus represents an effort to link disparate files and permit many different types of research. Cross-sectional linkage and longitudinal linkages are now relatively easy. In the past, life events (changes in family composition, mobility, and migration) made tracing individuals through time more difficult. In building the data base, we had to work with a family identification number that changes with some life events (becoming 18, divorce, remarriage). Under these circumstances, the linkage was difficult and time consuming. The costs involved in building this data base were large—perhaps on the order of $400,000 (Canadian dollars). On the other hand, the data base has supported approximately 160 published papers and is currently vital for a large research effort.

Generalized capability for linkage. The capability to perform record linkage as "files of opportunity" present themselves is very valuable. Such a capability enables a research group to link to:

1. Local statistical files that may have recently become available: Such linkages—typically involving chronic disease registries (such as those for cancer and for acquired immunodeficiency syndrome [AIDS]), birth defects, and mortality data—can provide the beginnings for a locally maintained comprehensive data base.

2. Survey data: Survey information linked with claims can provide a more complete picture of variables on both data sets. For example, in Manitoba, the relationships among functional status, self-reported health status, and utilization have been investigated through linking a multiwave panel study (Roos, N.P. and Havens, 1991; Roos, N.P., Montgomery, and Roos, L.L., 1987; Roos, N.P., Shapiro, and Tate, 1989).

3. Clinical data: Linking clinical data from one hospital's anesthesia quality assurance program with hospital claims and vital statistics information has produced a very rich Manitoba data set on preoperative status of patients and operative outcomes (Cohen and Duncan, 1988b; Roos, L.L., Roos, N.P., Fisher, and Bubolz, 1990).

This generalized capability for linkage may be increasingly important for research centers. As more hospital and physician records become computerized, linkages involving the medical record will become more frequent. Using clinical information for checking on administrative data is a significant component of research on medical treatment effectiveness. If primary data collection can be done with the possibilities for record linkage in mind, then long-term studies involving repeated interviews may be linked with longitudinal administrative data, multiplying the research opportunities many-fold.

Links with clinical investigations have proved important in increasing the plausibility of research that relies heavily on claims. Many physician-researchers have a deep-seated distrust of relying on claims data alone. A

recent study of the long-term outcomes of open and transurethral prostatectomies would probably not have been accepted for The *New England Journal of Medicine* without having prospectively collected clinical data to support the inferences from claims data (Roos, N.P., Wennberg, Malenka and others, 1989).

Aside from the actual study results, this linkage between claims and clinical research also permitted checking on the amount of extra information actually provided by the clinical data. Several types of analysis highlighted an important methodological point: claims often can generate case-mix adjustments almost as good or as good as those provided by expensive primary data collection (Roos, L.L., Sharp, and Cohen, 1991; Roos, N.P., 1989). Growth in knowledge regarding how to use claims appropriately may further improve our ability to predict important outcomes from claims (Green, Passman, and Wintfeld, 1991; Romano and Luft, in this volume).

Linkage Approaches

Complete, high-quality record linkage may not always be possible. Because the most powerful identifiers (unique identification numbers and names) may well be absent, knowing whether a linkage is feasible can help prevent wasted effort. Not enough information may be available in both of the data sets. Related concerns involve strategy and efficiency; if the linkage of two files is possible, how should it be carried out? Often, a substantial portion of the records may be linked with a high probability of accuracy; then additional data collection can focus on resolving the remainder of the cases.

As discussed in L.L. Roos and Wajda (1991), progress is being made in estimating the information needed to (a) identify each person uniquely by a set of values of the linkage variables and (b) carry out the linkage in the face of error in these variables. Calculating the amount of information present in each of the two files being linked aids in choosing an appropriate linkage strategy. Complicated probabilistic techniques have been particularly important for searching mortality files in occupational health research when the probability of a given individual being found is unknown (Shannon, Jamieson, Walsh, and others, 1989). In other situations, the probability of a match between two data sets will be quite high; relatively little error may be present in the variables used for linkage. We have been working on guidelines for situations where simpler techniques can be applied; as noted in Newcombe, Fair, and Lalonde (1989, Pts. I and II), accuracy versus simplicity will often be an important tradeoff.

Deterministic matching—a simple approach. As described by L.L. Roos and Wajda (1991), deterministic matching generates links based on the number of agreements among individual identifiers on the two files. Candidate pairs of records (one from each file) are selected on the basis of the number of such variables that do agree. This tactic simplifies the record linkage methodology, making it more practical for quick jobs and for rapid processing on even small (that is, personal) computers. Choices must be made among different methods for dealing with cases that do not match exactly. For data of relatively high quality, deterministic matches may be suitable. When data sets with many coding errors or with relatively few matching variables are linked, however, probabilistic methods will likely be more cost-effective. In addition, if many ties are generated, probabilistic matching will have to be used for the linkage. Approaches may also be combined; deterministic techniques may be used as a first step, producing files that can be analyzed to generate probabilistic weights for further matching. L.L. Roos and Wajda (1991) and Wajda, Roos, Layefsky, and Singleton (1991) provided some specific examples of deterministic matching.

Probabilistic matching—more complex approaches. Smith (1984) discussed probabilistic matching as follows:

> Agreements of various identifying items will generally argue in favor of a linkage, whereas disagreements will argue that the records relate to different people. Numerical weights can be used to quantify the fact that rare names, rare birthplaces, and such, carry more discriminating power when they agree than do their common counterparts.

Newcombe (1988) used a FREQUENCY/RATIO formula to express the basic principle of probabilistic linkage:

$$\text{FREQUENCY} = \frac{\text{frequency of outcome (x, y) among LINKED pairs}}{\text{frequency of outcome (x, y) among UNLINKABLE pairs}}$$

$$\text{RATIO}$$

where:

- x indicates the identifier and its value on the record from the file initiating the search (record A):
- y indicates the identifier and its value on the record from the file initiating the search (record B):

LINKED pairs may refer either to all linked pairs, or to a defined subset. *UNLINKABLE* pairs may refer either to all unlinkable pairs, or to a defined subset, provided the linked and the unlinkable sets (or subsets) are otherwise strictly comparable with each other.

Logarithms are used in this approach "...so the weights can be added and a total weight is achieved at the end indicating the probability that two records do or do not match" (Smith, 1984). On the logarithmic scale, the numerical weight is an estimate of the odds that the two records under consideration do, in fact, refer to the same individual—that is, they should be linked— as opposed to referring to different individuals—that is, they should not be linked (Howe and Lindsay, 1981). Zero represents odds of 1:1 that the linkage is a correct one, each added unit doubling the odds and each subtracted unit halving them. For example, +1 and +2 represent odds of 2:1 and 4:1, respectively, in favor of a correct match, whereas -1 and -2 represent odds of 1:2 and 1:4 and so argue against a correct match (Newcombe, Smith, Howe, and others, 1983). The weights represent potential links in decreasing order of certainty and are used to distinguish among such links (Howe and Lindsay, 1981; Newcombe, Smith, Howe, and others, 1983).

There are two types of probabilistic matching that differ according to how weights are generated. A linkage with nonspecific weights generates such weights based on whether a given identifier agrees or disagrees with its counterpart on the other file. This approach will also resolve more multiple matches (a record on one file linking with two or more records on another file) than will deterministic matching. The weights are relatively easy to calculate and may be satisfactory for many linkage projects.

Probabilistic linkage with outcome-specific weights both resolves matches more accurately and involves more difficult weight calculation. Each outcome among several is assigned a weight. For example, one set of outcome-specific weights is "based on the numbers of agreements when the searches are initiated by records carrying the particular letter as a first initial (of a name)" (Newcombe and Abbatt, 1983). Weights using year or month of birth (or death) can be calculated to take partial agreement (such as a discrepancy of 1 month or year) into account. This method is particularly valuable for searching huge mortality data bases when false positives are likely and when it is not known whether an individual is alive.

Manitoba experiments with deterministic linkage found very few false positives (records that link when they should not) accompanying conservative use of the technique (Roos, L.L. and Wajda, 1991). Such findings support the appropriateness of this simpler methodology, even though the weights for each linkage are not estimated. As suggested by the figure, probabilistic matching techniques, through the use of weights, provide an estimate of the probability of a given link representing information from the same individual (Dale, 1989). "Grey areas" where links may or may not be correct can thus be identified through probabilistic matching.

Estimating information. The amount of information available for record linkage may substantially exceed the minimum necessary; typically, in such information-rich situations, either a large number of variables or several particularly powerful ones (name, address, social insurance number) are available on both files. As discussed earlier, a relatively simple deterministic approach may be appropriate here. Because the simpler approaches to record linkage do not use all the information potentially available, they will be inappropriate in information-poor settings. In information-poor linkages, when relatively few variables are available for linkage, the analytical problem is one of distinguishing among two or more tied cases. These are cases that may be resolved by more sophisticated linkage methods (specific probabilistic matching) or by additional data.

Our simulations have shown that specific probabilistic matching typically produces only a 2-3 percent advantage over the nonspecific probabilistic methods (which have recently been automated). This is very useful in raising the percentage linked from 93 percent to 96 percent but does not help much when only 80 percent of the individuals who should be linked are linkable by nonspecific probabilistic methods.

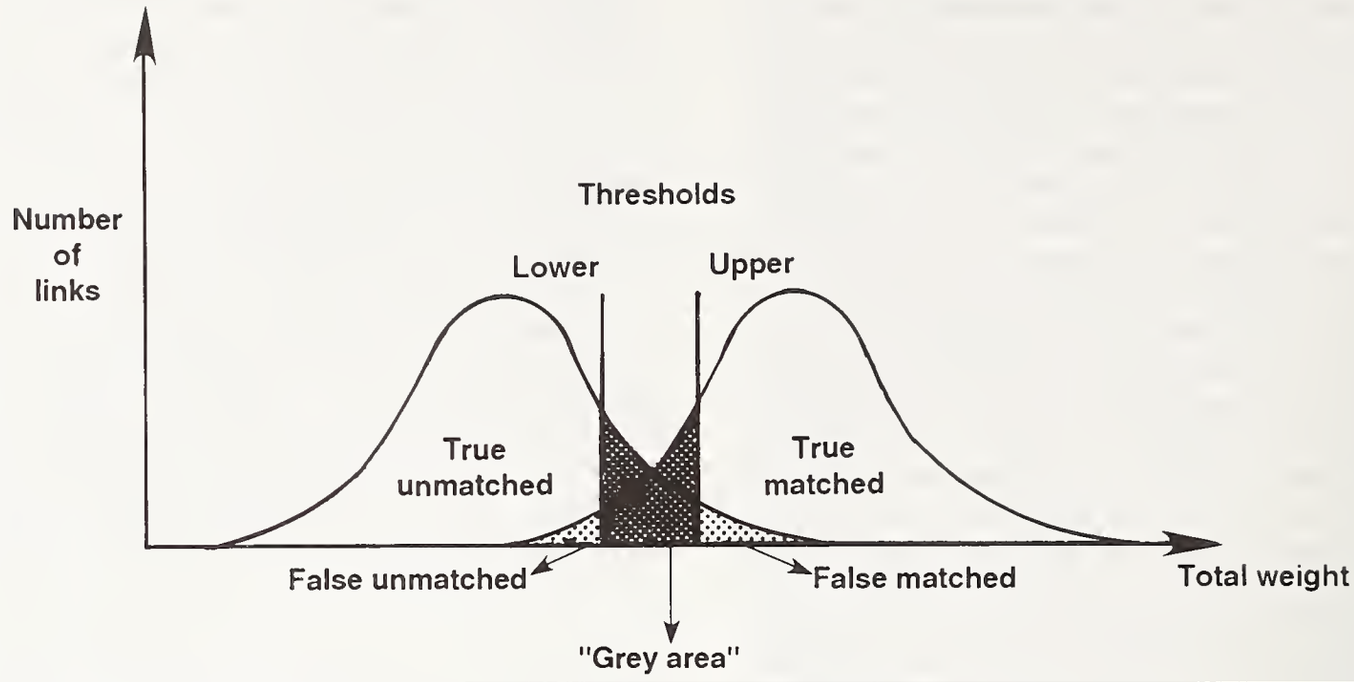
In some circumstances, record linkage may prove infeasible without the incorporation of additional identifiers into one or more data sets. Additional data may be generated by going over the files to try to find an additional variable or two to aid the linkage or by primary data collection. Adding one more variable is likely to resolve most of the tied cases; expensive primary data collection (perhaps adding initials, names, or identifying numbers) might be restricted to the tied cases.

A certain minimum amount of information is required to identify each person as unique. The minimum number of variables that have to agree to permit linkage can be estimated; almost every individual on each file must fit into a "box" defined by a unique combination of identifiers. Because of measurement error, however, an extra variable or (more likely) variables will be necessary, particularly if the simpler linkage techniques are used. L.L. Roos and Wajda (1991) discussed how to estimate the information needed for linkage.

Discussion

Registries are especially likely to become greater users of record linkage techniques. Because of their experience with reconciling data from different sources, cancer researchers seem particularly likely to be interested in linking various files. Interest in record linkage at other chronic disease registries (such as those for AIDS)

Figure. Use of probabilistic matching



Source: Dale, 1989.

will almost certainly grow over the next few years. The ability of linkage software to run on personal computers may be particularly important for registries where maintaining confidentiality is of the highest priority (such as AIDS registries). Linkages that can be run on a computer in a single individual's office bypass any security problems associated with mainframes and larger computer centers. Finally, because of their increasing attention to large data bases and postmarketing surveillance, workers in the growing field of pharmacoepidemiology seem likely to be attracted to record linkage techniques (Rawson and D'Arcy, 1991).

Health services researchers are only beginning to explore the potential of comprehensive data bases. The deterministic linkages (using social security number) and mortality checks coordinated by the Health Care Financing Administration in the United States have led to a great deal of work. Persuading States to collect and provide identifiers to carefully screened projects and investigators would greatly facilitate research, as indicated in a recent report to Congress prepared by the Agency for Health Care Policy and Research (AHCPR) (AHCPR, 1991). Links to clinical data bases are likely to necessitate more sophisticated linkage techniques and software.

Luft (Personal communication, May 1991) suggested several specific linkage applications in medical effectiveness research. As noted in the AHCPR report to Congress (1991), linkage of State health care utilization data files presents many possibilities. Some large, population-based hospital data sets (such as those for California) do not include patient name or identifying number when released to academic researchers. If enough identifying information was available (such as birth date, sex, place of residence, marital status, diagnoses and procedures, dates of admission and discharge, discharge status, hospital identifier), tracking patient transfers through probabilistic linkage would be feasible, even in large States. Moreover, because administrative data generally include fewer diagnoses than are likely to be noted by clinician/researchers, having additional diagnostic information is very valuable. Data from the same individual before an index hospitalization could be used to increase the information available on comorbidity.

Such linkage to build longitudinal histories within a single file is only one example. Because California hospital discharge abstracts contain a great deal of information on diagnoses and procedures, linkage of California patients over 65 to Medicare data would provide extra data on comorbidity for studying outcomes. Such linkages would help deal with questions concerning the appropriate use of several measures of mortality—such as inpatient, 30 days, 90 days, and so on—since only inpatient mortality is available on the California discharge abstracts.

Some tension may develop between national mortality data base and local (provincial or State) registry ap-

proaches. Considerable cost savings accrue, at least in the Canadian provinces where we have worked, if in-house followup is feasible for the research task at hand. Differences between the data resources are summarized below:

Local Database and Registry	National Mortality Data Base
Migration between provinces or states and migration out of country handled as loss to followup	Interprovincial migration does not cause loss to followup
	Does not identify migration out of country
Coordinated with other data files (such as hospital, medical)	Used for death clearance

One salient issue in record linkage concerns the extent to which linkage to State or provincial mortality data can substitute for more expensive death clearance with a central registry. Statistics Canada reported 4 percent interprovincial migration between the 1981 and 1986 provincial censuses (Fair, 1990). Our Manitoba experience has indicated a higher migration rate in young adults and a lower rate in older adults who are likely to be both the victims of chronic disease and candidates for surgery. Thus, the prospects for long-term followup by State and provincial registries will vary according to the age of the individuals being studied. The decision as to local or central followup will also be dependent on cost tradeoffs and other goals.

If tracing morbidity, as well as mortality, is a research goal, it might be necessary to direct resources toward organizing hospital and medical claims for the study of such morbidity. The comprehensive data base is enriched by each linkable file. Many papers other than those originally planned become possible, and the data can be a resource for a number of investigators.

Another important issue concerns facilitating record linkage within different research settings; here, flexibility of the strategy for matching is very important (Wajda and Roos, L.L., 1987). Recent developments in software (fourth-generation languages and portable, modularized routines) permit using the computer in a rapid feedback, "experimental" mode on a personal computer or workstation. A good strategy using versatile software can provide results as high in quality as those available from large systems such as GRLS (Generalized Record Linkage System) (Newcombe, 1988; Shannon, Jamieson, Walsh, and others, 1989), and CAMLIS (California Automated Mortality Linkage System) (Arellano, Petersen, Petitti, and Smith, 1984) at a fraction of the cost. Record linkage need not be confined to longitudinal research but can expand the data sets available for work in health services, occupational health, and epidemiology (both clinical and population-based). Such data promise to be increasingly relevant for medical informatics (Greenes and Shortliffe, 1990) generally and the medical effectiveness research program specifically.

References

Acheson, E.D. (1967). *Medical record linkage.* London: Oxford University Press.

Acheson, E.D. (1968). Introduction. In Acheson, E.D. (Ed.), *Record linkage in medicine. Proceedings of the international symposium, Oxford, U.K.; July 1967.* Edinburgh: E. & S. Livingstone.

Acheson, E.D. (1987). In Baldwin, J.A., Acheson, E.D., and Graham, W.J. (Eds.), *A textbook of medical record linkage.* Oxford: Oxford University Press.

Agency for Health Care Policy and Research. (1991). *Report to Congress: The feasibility of linking research-related data bases to Federal and non-Federal medical administrative data bases* (AHCPR Pub. No. 91-0003). Rockville, MD: Author.

Arellano, M.G., Petersen, G.R., Petitti, D.B., and Smith, R.E. (1984). The California Automated Mortality Linkage System (CAMLIS). *American Journal of Public Health,* 74, 1324-1330.

Buehler, J.W., Berkelman, R.L., and Curran, J.W. (1989). Reporting of AIDS: Tracking HIV morbidity and mortality. *Journal of the American Medical Association,* 262, 2896-2897.

Carpenter, M. and Fair, M.E. (1989). Development of a data collection package for long term medical follow-up in Canada. In Carpenter, M. and Fair, M.E. (Eds.), *Proceedings of the record linkage sessions and workshop, Canadian Epidemiology Research Conference, Ottawa, August 30-31, 1989* (pp. 223-238). Ottawa: Statistics Canada.

Case, D. and McGregor, S. (1983). Large system implementation in SAS: The resource management system using SAS, SAS/FSP, SAS/Graph and the Merrill database. In *SUGI Proceedings* (pp. 23-26). Cary, NC: SAS Institute.

Cohen, M.M. and Duncan, P.G. (1988a). Physical status score and trends in anesthetic complications. *Journal of Clinical Epidemiology,* 41, 83-90.

Cohen, M.M. and Duncan, P.G. (1988b). Postoperative follow-up and quality of care. *Seminars in Anesthesia,* 7, 270-277.

Dale, D. (1989). Linkage as a part of a production system: The Ontario Cancer Registry. In Carpenter, M. and Fair, M.E. (Eds.), *Proceedings of the record linkage sessions and workshop, Canadian Epidemiology Research Conference, Ottawa, August 30-31, 1989* (pp. 193-200). Ottawa: Statistics Canada.

Fair, M.E. (1990). *Looking for delayed harm in a mobile population in the 1990s.* Unpublished manuscript.

Fair, M.E. and Lalonde, P. (1991). Application of exact ODDS for partial agreement of names in record linkage. *Computers and Biomedical Research,* 24, 58-71.

Godber, G. (1968). Record linkage. In Acheson, E.D. (Ed.), *Record linkage in medicine. Proceedings of the international symposium, Oxford, U.K.; July 1967* (pp. 1-4). Edinburgh: E. & S. Livingstone.

Green, J., Passman, L.J., and Wintfeld, N. (1991). Analyzing hospital mortality: The consequences of diversity in patient mix. *Journal of the American Medical Association,* 265, 1849-1853.

Greenes, R.A. and Shortliffe, E.H. (1990). Medical informatics: An emerging academic discipline and institutional priority. *Journal of the American Medical Association,* 263, 1114-1120.

Howe, G.R. (1989). The use of computerized record linkage in health studies in Canada: The past, present, and future. In Carpenter, M. and Fair, M.E. (Eds.), *Proceedings of the record linkage sessions and workshop, Canadian Epidemiology Research Conference, Ottawa, August 30-31, 1989* (pp. 3-18). Ottawa: Statistics Canada.

Howe, G.R. and Lindsay, J. (1981). A generalized iterative record linkage computer system for use in medical follow-up studies. *Computers in Biomedical Research,* 14, 327-340.

Newcombe, H.B. (1957). Detection of genetic trends in public health. In *Effect of radiation on human heredity.* Geneva: World Health Organization.

Newcombe, H.B. (1968). Early stages of linked records. In Acheson, E.D. (Ed.), *Record linkage in medicine. Proceedings of the international symposium, Oxford, U.K.; July 1967* (pp. 7-34). Edinburgh: E. & S. Livingstone.

Newcombe, H.B. (1988). *Handbook of record linkage.* New York: Oxford University Press.

Newcombe, H.B. and Abbatt, J.D. (1983). *Probabilistic record linkage in epidemiology: Principles employed.* Ottawa: Eldorado Resources, Ltd.

Newcombe, H.B., Fair, M.E., and Lalonde, P. (1989). Discriminating powers of partial agreements of names for linking personal records. Part I: The logical basis and Part II: The empirical test. *Methods of Information in Medicine,* 28, 86-91, 92-96.

Newcombe, H.B., Smith, M.E., Howe, G.R., and others. (1983). Reliability of computerized versus manual death searches in a study of the health of Eldorado uranium workers. *Computers in Biology and Medicine,* 13(3), 157-169.

Rawson, N.S.B. and D'Arcy, C. (1991). "Validity" and reliability: Idealism and reality in the use of computerized health care databases for pharmacoepidemiological research. *Post-Marketing Surveillance,* 5(1), 31-55.

Roos, L.L., Nicol, J.P., Johnson, C., and Roos, N.P. (1979). Using administrative data banks for research and evaluation: A case study. *Evaluation Quarterly,* 3, 236-255.

Roos, L.L., Nicol, J.P., and Wajda, A. (1985). Improving the quality of data banks through linkage. *Chronic Diseases in Canada,* 5, 81-82.

Roos, L.L., Roos, N.P., Cageorge, S.M., and Nicol, J.P. (1982). How good are the data? Reliability of one health care data bank. *Medical Care,* 20, 266-276.

Roos, L.L., Roos, N.P., Fisher, E.S., and Bubolz, T.A. (1990). Strengths and weaknesses of health insurance data systems for assessing outcomes. In Gelijns, A.C. (Ed.), *Medical innovation at the crossroads. Volume I. Modern methods of clinical investigation* (pp. 47-67). Washington: National Academy Press.

Roos, L.L., Sharp, S.M., and Cohen, M.M. (1991). Comparing clinical information with claims data: Some similarities and differences. *Journal of Clinical Epidemiology,* 44, 881-888.

Roos, L.L., Sharp, S.M., and Wajda, A. (1989). Assessing data quality: A computerized approach. *Social Science and Medicine,* 28, 175-182.

Roos, L.L. and Wajda, A. (1991). Record linkage strategies: Part I. Estimating information and evaluating approaches. *Methods of Information in Medicine,* 30(2), 117-123.

Roos, N.P. (1989). Using administrative data to study outcomes: Developing control groups and adjusting for case severity. *Social Science and Medicine,* 28, 109-113.

Roos, N.P. and Havens, B.J. (1991). Predictors of successful aging: A twelve year study of Manitoba elderly. *American Journal of Public Health,* 81, 63-68.

Roos, N.P., Montgomery, P., and Roos, L.L. (1987). Health care utilization in the years prior to death. *Milbank Quarterly,* 65, 231-254.

Roos, N.P., Shapiro, E., and Tate, R.B. (1989). Does a small minority of elderly account for a majority of health care expenditures? A sixteen year perspective. *Milbank Quarterly,* 67, 347-369.

Roos, N.P., Wennberg, J.E., Malenka, D.J., and others. (1989). Mortality and reoperation after open and transurethral resection of the prostate for benign prostatic hyperplasia. *New England Journal of Medicine,* 320, 1120-1124.

SAS Institute. (1985). *SAS introductory guide.* Cary, NC: Author.

Shannon, H.S., Jamieson, E., Walsh, C., and others. (1989). Comparison of individual follow-up and computerized record linkage using the Canadian Mortality Data Base. *Canadian Journal of Public Health,* 80, 54-57.

Smith, M.E. (1984). Record linkage: Present status and methodology. *Journal of Clinical Computing*, 13, 52-69.

Wajda, A. and Roos, L.L. (1987). Simplifying record linkage: Software and strategy. *Computers in Biology and Medicine*, 17, 239-248.

Wajda, A., Roos, L.L., Layefsky, M., and Singleton, J.A. (1991). Record linkage strategies: Part II. Portable software and deterministic matching. *Methods of Information in Medicine*, 30(3), 210-214.

Young, T.K., Roos, N.P., and Hammarstrand, K.M. (1991). Estimated burden of diabetes mellitus in Manitoba according to health insurance claims: A pilot study. *Canadian Medical Association Journal*, 144(3), 318-324.

Appendix: Software

As noted by Wajda and Roos (1987), Manitoba record linkage work has built on a fourth-generation language, SAS (Statistical Analysis System), with accompanying macroprocessor, to develop a user-friendly system for both exact and probabilistic matching. This system, in conjunction with the tactics outlined here, has made the linkage process markedly faster in our ongoing research.

The use of macros in a fourth-generation language fits well with a modular approach that:

1. Permits both independent execution of modules and chaining modules together; this can help better organize computer resources (for example, to meet computer center restrictions on maximum run length);
2. Allows examination of intermediate output;
3. Provides better control of the programming process, in that results from each module can be debugged and discussed separately;
4. Has the advantage of being able to draw on capabilities already available in SAS; and
5. Facilitates learning by implementing a consistent command structure across the modules; users who already know SAS find the package especially easy to learn.

Distinguishing between a program to carry out record linkage and a system that, among other tasks, will perform record linkage is worthwhile. Desirable system capabilities include those for data management, reporting, and statistical analysis. Fourth-generation languages such as SAS may represent a critical technology for dealing with large data bases (Case and McGregor, 1983).

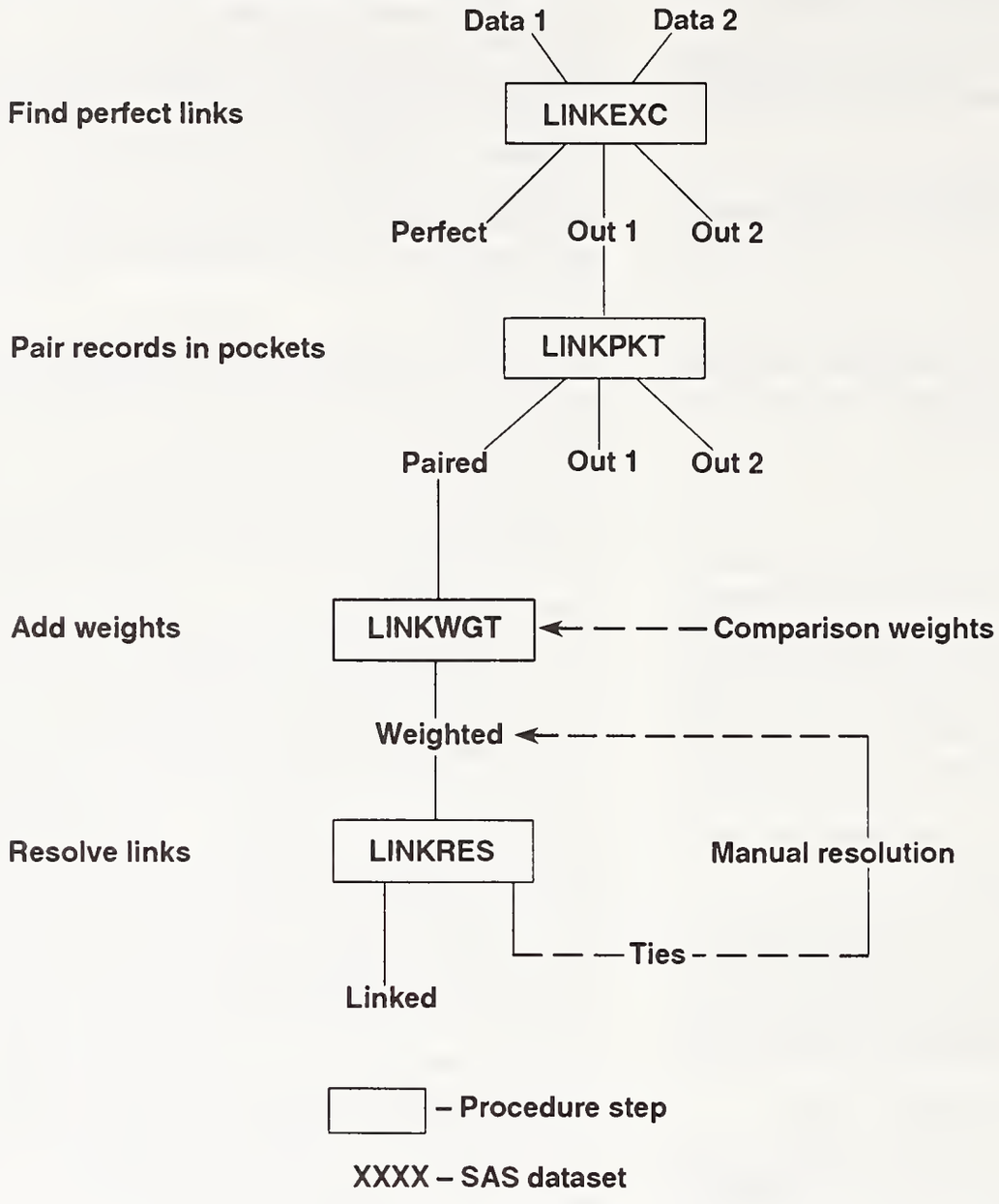
LINKS is a series of modules designed to perform data management and record linkage functions. The modules are SAS macros that execute the programs using such user-defined parameters as data set names, lists of variables, and comparison rules. The syntax of LINKS is almost identical to other SAS procedures, which makes it very easy for people familiar with SAS. Both exact and probabilistic matching are carried out, building on the methodology developed at Statistics Canada (Smith, 1984).

The first two modules, _TESTPWR and _TESTPKT, are used to generate information about the structure of each data set, the importance of each variable, and the feasibility of linkage. A data set is any group of records from a single source with a common format. A variable is a set of data values for the same measurement—an individual's sex, year of birth, and so forth (SAS Institute, 1985). This preprocessing is extremely important in reducing overall costs. Alternate recoding systems can be studied and may significantly improve the quality of the linkage. Specifically, _TESTPWR prescreens the data sets to indicate which variables are likely to be most useful. An output table presents frequency distributions for all variables, the number of missing values, and two measures of the relative power of each variable in the linkage. Details on these measures—discriminating power and Shannon entropy—are provided for each person. The procedure splits and sorts the data sets into the "pockets" containing specific values of each variable. An output table lists the variables that define each pocket, the number of pockets, and the minimum, maximum, and mean number of records in each pocket.

As presented in the Appendix Figure, the four modules to carry out the record linkage are structured to permit easy changes in variables, algorithms, and weights employed. The system allows the user to run either deterministic linkage (by running a series of _LINKEXC programs) or probabilistic models (using _LINKPKT, _LINKWGT, and _LINKRES), as done at Statistics Canada. _LINKEXC compares two data sets for records that agree on all comparison variables—that is, it finds all the "perfect" matches. These matches are saved to an output file. Remaining unmatched records are also saved in their own output files. _LINKEXC reduces the number of records that need to be considered for probabilistic matching, providing major savings in computer resources.

The probabilistic matching builds on the Statistics Canada strategy for dealing with the large number of conceivable matches for any record by selecting "an item or items to serve as pocket identifiers to sequence the files. This results in a comparison space of small blocks in which all records are compared" (Smith,

Appendix Figure. Linkage procedures flowchart



Source: Wajda and Roos, L.L., 1987.

1984). This process brings together potentially linkable pairs of records into pockets, reducing the amount of computing necessary. The first phase of this probabilistic matching, _LINKPKT, places records within pockets if they agree on a minimum number of user-specified variables. Again, three output files are saved: one with the new paired records, a second with the records remaining from the first data set, and a third with records remaining from the second data set.

_LINKWGT then attaches probabilistic weights based on comparison of variables to the records paired by _LINKPKT. The user provides the variables, the weighting factors, and any other conditions on the weights. The procedure _LINKRES resolves the links using the probabilistic weights generated by _LINKWGT. Two output files are produced: one with the resolved links and the other with the unresolved links.

Flexibility in managing pocket size is particularly important; probabilistic matching within each pocket is the most computer-intensive part of the package. _LINKPKT facilitates reducing the number of potentially linkable pairs by incorporating stricter criteria for entry into the pocket (for example, agreement on four variables rather than three); this may save substantial computer time with few changes in the records eventual-

ly linked. Because the modules can be arranged in any order, a probabilistic match involving small pockets (specifying, for example, that records must agree on five variables) could be followed by a probabilistic match involving cases agreeing on four but not five variables.

A Discussion of Record Linkage

Martha E. Fair

The paper by Roos, Wajda, Nicol, and Roberts (in this volume) has very succinctly described the history, methods, software, and uses of record linkage. A number of the historical details may require further clarification as to early concepts of record linkage. I would like to emphasize the importance of a broader view of linkage both in terms of definition and in complementary execution at regional, national, and international levels. I will offer some comments on linkage approaches, defining roles in order to promote collaboration and to avoid duplication of efforts; elaborate on some possible ways to reduce lost-to-followup for epidemiology studies; highlight further examples of uses of record linkage in Canada, particularly with respect to medical effectiveness research methods; and talk about software and some features to consider in terms of flexibility and cost.

Early Concepts of Record Linkage

The early concepts of record linkage were very broad. They included a very wide range of files and envisaged efficient administrative and statistical uses of data at local, state, and national levels. In the original paper "Record Linkage," Dunn (1946) simply used the term "record linkage" not "medical record linkage." This is important because his concept was much broader than just health needs as discussed in the paper by Roos and others. What he envisaged was linkage of the "records of the principal events of life" for each individual into a "book of life." He pointed out that Canada is in an ideal position to exploit linkage of many kinds, using a "life records index" to show where the important life records of an individual are kept. He envisaged that numerous official organizations at national, state, and local levels would rely heavily on certain chapters in the life records index. He talked about binding the books of life into volumes. The index would indicate where all the most important records of an individual could be found. He applied the record linkage principle to many social and economic services and benefits so that they might be better directed to the real needs of people. Some of the administrative needs at the state level that he mentioned were for certification, legal proof of identity, title to property, dead/alive status, and death clearance of many active files.

In 1959, Canadian (not British) researchers applied the computer to the problem of sorting and matching personal records with discrepant identifying data (Newcombe, Kennedy, Axford, and James, 1959). Development of the theory in a more rigorous mathematical fashion was carried out in 1969 at Statistics Canada (Fellegi and Sunter, 1969). I point this out because many of the basic principles still being used today were described in these early papers. The Oxford linkage study in Britain was set up in 1962 and adapted the Canadian methods to link hospital and death records into personal histories. In the Introduction to the *Textbook on Medical Record Linkage* (Baldwin, Acheson, and Graham, 1987), Acheson recounts the early progress of the Oxford record linkage study:

> Initially we had hoped to link the records into a series of person cumulative files by using the National Health Service number but it soon became clear that this number was only available on a minority of records. Fortunately Dr. Alex Barr, statistician to the Oxford Regional Hospital Board (as it was then), had heard Dr. Howard Newcombe expound his elegant method of linking records, in spite of omissions and discrepancies in the data, by using names, dates of birth, and other information (Newcombe, Kennedy, Axford, and James 1959). We adopted a system based on similar principles, but modified to the British scene in 1973. This has subsequently been much improved and refined. (p. 2)

Linkage Approaches

In the discussion on linkage approaches, reference is made to Dale's (1990) figure 1. This figure creates the

Mrs. Fair is Chief, Occupational and Environmental Health Research Section, Canadian Centre for Health Information, Statistics Canada, Ottawa, Ontario.

misleading impression that the matched and unmatched pairs are about equal in number. Much confusion has resulted in the past because people have overlooked the great preponderance of the false links.

The process of separating out the true matches is, in reality, a stepwise elimination of the false ones. It is not so much a matter of picking needles out of a haystack, as it is one of progressively getting rid of the haystack without losing the needles (Newcombe, 1988).

The most important role played by identifying information is not singling out the good matches, but rather identifying the bad ones, which is achieved by weighting the size of the files being linked. If we were to start by ignoring all the identifiers being used in a match, the potential number of candidate pairs for linkage would be the product of the number of records in the two files that are being linked. As increasing amounts of identifying information are brought into play, a stepwise narrowing of the remaining candidate pairs occurs through the successive elimination of "failed pairs." The problem with uniquely identifying an individual in a file of 20 million records is more difficult than identifying the person in a file of 100,000 because one has to discard from consideration far more unlinkable pairs. When the information needed for linkage is estimated, the file size must be taken into account. Whether to use the deterministic or probabilistic linkage approach depends very much on the quality of the data being linked, the time period, and the accuracy required.

Communication and Collaboration

It is important that a communications network and agreements be set up to define the roles and responsibilities of various organizations at the provincial/state and national/federal levels. Roos and colleagues (in this volume) have interjected this challenge into their discussion by stating that, "Considerable tension may develop between national mortality data base and local (provincial or state) registry approaches."

In Canada, neither the national or provincial authorities or the researchers have yet made more than a modest start at tapping the immense resources at their disposal for health research. Each level of participation should develop its own resources so that the contributions complement each other rather than compete. There will be instances where considerable cost savings can accrue from doing work in house, if this is feasible. In general, the closer in time and place to the origin of the source documents, the easier the task.

Statistics Canada has tried to encourage and facilitate provincial and local agencies carrying out linkages by sponsoring workshops and training wherever possible within our limited budget (Carpenter and Fair, 1990; Coombs and Singh, 1987; Howe and Spasoff, 1986). Such workshops include a record linkage workshop (November 15, 1991) following a methodology symposium on *Spatial Issues in Statistics* and another (July 1992) associated with the International Association of Cancer Registries annual meeting in Ottawa. In the United States, further international stimulus and communication were given by conferences organized by Fritz Scheuren at the Internal Revenue Service (for example, Kilss and Alvey, 1985). Efforts to study undercoverage in the 1990 Decennial Census have advanced work in this area (for instance, Winkler, 1991).

In the United States, the approach taken in many mortality studies has been to do a linkage with the National Death Index (NDI) which started in 1979 (National Center for Health Statistics, 1990). The user is responsible for assessing the results of the NDI search and determining which NDI records are true matches, questionable matches, and false matches. The researcher then must write to the appropriate State vital statistics offices for copies of relevant death certificates, code the certificates, and analyze the files.

In Canada, we are fortunate to have cause of death coded on the Canadian Mortality Data Base file, and these files date back to 1950. We often get much more involved in the methodology of the linkage and the research project. In addition to mortality searches, the results from such studies have been coordinated with other files. For example, one can coordinate and link the results with other data sources such as the National Cancer Incidence Reporting System, individual exposure histories, and morbidity data sources (Smith and Newcombe, 1982).

Reducing Lost-to-Followup

Roos and his colleagues (in this volume) did not mention total vital status followup in connection with linkage as carried out in the provinces, although migration was referred to. Steps have been taken to carry out alive followup both at the provincial and at the national level to complement death followup for studies. For example, in a study of Ontario miners, provincial resources such as driver's license files are used to determine alive status. Other potential sources might include tax rolls, voter lists, free drug eligibility lists, and so forth. At the provincial level, a number of linkages are carried out to ensure the eligibility of persons for provincial health insurance purposes (Thibault, 1989). At the national level, income tax records are used to determine the alive status of individuals.

Uses of Record Linkage

In addition to some of the uses given by Roos and others, it may be helpful to highlight additional studies cur-

rently in progress that could be of interest to medical effectiveness researchers.

Breast cancer is a leading cause of disability and death from cancer in women and is the largest single cause of death in women aged 35-59. A National Breast Screening Study was undertaken in Canada as a randomized controlled trial organized by the National Cancer Institute of Canada, Epidemiology Unit, University of Toronto, to assess the efficacy of mammography and physical examination in screening for breast cancer (Day and Miller, 1988). A number of provinces across Canada are following up with early detection programs and dedicated screening centers.

The study recruited approximately 90,000 women aged 40-59 from across Canada to participate for a 5-year period from 1980. The main end point for assessing the results of screening is the mortality from breast cancer in the study groups compared with that in the controls.

The methodology developed to follow up this cohort consisted of a variety of techniques to determine vital status. In addition to the mortality search, an alive search was carried out using records extracted from the income tax files and the inclusion of alternate records following checks of name spellings in the study cohort. The analysis of the study is currently in progress, and particular interest is being placed on the 40- to 49-year-old groups.

Day and Miller (1988) outlined a number of important research issues. These included, for example, the comparative value of mammography and physical examination in different age groups and the value of screening women under 50 years of age.

Cardiovascular diseases (CVDs) are the leading causes of death, potential years of life lost, and utilization of health care services in Canada. Since the late 1960s, Canada and many other western countries have experienced a noticeable decline in mortality from CVDs and particularly from ischemic heart disease (IHD). Efforts are under way worldwide to study the determinants of change in cardiovascular mortality in prospective studies. The idea of linking routinely collected hospital separation records of IHD to Canadian mortality records was initiated and field tested in the provinces of Nova Scotia and Saskatchewan. A feasibility study relating to acute myocardial infarction (AMI) was initiated, and it used information from provincial morbidity files and the Canadian Mortality Data Base. A sample of AMI cases from the hospital admission data bases of Nova Scotia and Saskatchewan were field validated, and individual histories were created to differentiate recurrent from incident cases. The incidence for fatal and nonfatal AMIs was greater in Nova Scotia than in Saskatchewan over three time periods examined (1977, 1981, and 1985). The results of the early study were published elsewhere (Nova Scotia-Saskatchewan Cardiovascular Disease Epidemiology Group, 1989).

Extensive work in Canada has taken place at both the provincial and national levels using record linkage in building, maintaining, and using cancer registries (Clarke, 1986; Dale, 1990). The generalized record linkage system was used in creating the Ontario Cancer Registry. A National Cancer Incidence Reporting System also includes information on cancer for the complete Canadian population. This model could be used in creating other types of disease registries.

A number of radiation studies have been carried out in Canada. These data also served as input into international studies. A study on farmers is investigating the relationships between causes of death and various agricultural and sociodemographic variables. Of particular interest in this study are relationships between exposure resulting from various farm practices and cancers of sites that have been implicated in previous investigations of farmers and other related workers. The major files being used are the 1971 Censuses of Agriculture and Population as well as the Canadian Mortality Data Base.

Software

Roos and his coworkers (in this volume) advocate a linkage approach that is tied to the Statistical Analysis System (SAS). The description of the mechanics of the LINKS approach is useful, but it falls short of saying whether the flexibility is equal to that which would be possible when not tied to SAS. We certainly have used SAS extensively for related record linkage work in research projects and have found it helpful.

In regard to cost and flexibility, some of the tests required to make a genuine benchmark comparison between a variety of systems require more careful analysis. It is important to clarify whether one is really evaluating software or rather the quality of the data files that are being used in the linkage.

An important first question in looking at software is how general a system must be for the task at hand? At Statistics Canada, which is a centralized statistical agency, there is a need not only to carry out a linkage internal to files (for example, to create individual health histories) and for two-file linkages (for example, linking an occupational cohort file to mortality), but also the flexibility to carry out linkages of entities (such as farms and businesses) and to carry out quality assessments (such as for reverse record checks for the census to determine undercoverage).

Other questions that may need to be considered in evaluating software for health research applications include:

- Is there a limit on the size of projects that should be carried on a personal computer?
- When cost is being determined, what factors are being considered? Should one compare the actual computer costs or the total project costs?
- Are administrative and other related costs, such as standardizing formats, editing of data, manual resolution, recoding of disease codes, and creation of analysis files, taken into account? In our experience, we find that the costs of a research project often are increased by the lack of good quality data files.
- How experienced is the staff using the data sources and linkage methodology?
- How flexible are the mapping features in linkage software? Can one generate and select whether there is a one-to-one, one-to-many, many-to-one, or many-to-many output?
- What about conflict report generation (for example, two individuals linking to one death)?
- What is the availability and flexibility of weight generation? It is not clear that nonlinks are generated to facilitate and ease the calculation of appropriate weights for items where the rules are complex, for example, place of work versus place of death or partial agreements of forenames.

No mention or reference is made to a new system that is currently being developed at Statistics Canada (Hill, 1990) which tries to address most of the above features. Considerable record linkage work is also being carried out at the U.S. Bureau of Census (Scheuren and Winkler 1991; Winkler, 1991).

Conclusion

Roos and his colleagues (in this volume) have successfully demonstrated that record linkage is an important research tool. Multiple data sources can be combined successfully using existing software to facilitate studies on medical effectiveness.

References

Baldwin, J.A., Acheson, E.D., and Graham, W.J. (Eds.). (1987). *Textbook of medical record linkage.* New York: Oxford University Press.

Carpenter, H. and Fair, M.E. (Eds.). (1990). *Proceedings of the record linkage sessions and workshop, Canadian Epidemiology Research Conference, Ottawa, August 30-31, 1989.* Ottawa: Statistics Canada.

Clarke, E.A. (1986). The Ontario Cancer Registry: From administration to research using GRLS. In Howe, G.R. and Spasoff, R.A. (Eds.), *Proceedings of the Workshop on Computerized Record Linkage in Health Research, May 21-23, 1986.* Toronto: University of Toronto Press.

Coombs, J.W. and Singh, M.P. (Eds.). (1988). *Statistical uses of administrative data: An international symposium*, November 23-25, 1987. Ottawa: Statistics Canada.

Dale, D. (1990). Linkage as a part of a production system: The Ontario Cancer Registry. In Carpenter, M. and Fair, M.E. (Eds.), *Proceedings of the record linkage sessions and workshop, Canadian Epidemiology Research Conference*, Ottawa, August 30-31, 1989. Ottawa: Statistics Canada.

Day, N.E. and Miller, A.B. (Eds.). (1988). *Screening for breast cancer.* Toronto: Hans Huber Publishers.

Dunn, H. (1946). Record linkage. *American Journal of Public Health*, 36, 1412-1416.

Fellegi, I.P. and Sunter, A.B. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Hill, T. (1990). GRLS-Future directions in record linkage. In Carpenter, M. and Fair, M.E. (Eds.), *Proceedings of the record linkage sessions and workshop, Canadian Epidemiology Research Conference, Ottawa, August 30-31, 1989.* Ottawa: Statistics Canada, 161-178.

Howe, G.R. and Spasoff, R.A. (Eds.). (1986). *Proceedings of the Workshop on Computerized Record Linkage in Health Research.* Toronto: University of Toronto Press.

Jordan-Simpson, D.A., Fair, M.E., and Poliquin, C. (1990). Canadian Farm Operators Study: Methodology. *Statistics Canada Health Reports*, 2(2), 141-155.

Kilss, B. and Alvey, W. (1985). Record linkage techniques - 1985. *Proceedings of the Workshop on Exact Matching Methodologies, Arlington, VA, May 9-10, 1985.* Washington: Internal Revenue Service.

Miller, A.B. (1988). The Canadian National Breast Screening Study. In Day, N.E. and Miller, A.B. (Eds.), *Screening for breast cancer* (pp. 51-58). Toronto: Hans Huber Publishers.

National Center for Health Statistics. (1990). *National Death Index users manual* (DHSS Pub. No. (PHS) 90-1148). Hyattsville, MD: U.S. Department of Health and Human Services.

Newcombe, H.B. (1988). *Handbook of record linkage - Methods for health and statistical studies, administration and business.* New York: Oxford University Press.

Newcombe, H.B., Kennedy, J.M., Axford, S.J., and James, A.P. (1959). Automatic linkage linkage of vital records. *Science*, 130, 954-959.

Nova Scotia-Saskatchewan Cardiovascular Disease Epidemiology Group. (1989). Estimation of the incidence of acute

myocardial infarction using record linkage: A feasibility study in Nova Scotia and Saskatchewan. *Canadian Journal of Public Health*, 80, 412-417.

Scheuren, F. and Winkler, W.E. (1991). Regression analysis of data files that are computer matched. *Proceedings of 1991 Census Annual Research Conference* (pp. 669-687). Washington: Bureau of the Census.

Smith, M.E. and Newcombe, H.B. (1980). Automated follow-up facilities in Canada for monitoring delayed health effects. *American Journal of Public Health*, 70, 1261-1268.

Thibault, N. (1989). L'effectif de la population du Quebec en 1986. Une comparaison entre le recensement et le fichier de l'assurance-maladie. *Cahiers quebecois de demographie*, 18, 323-341.

Winkler, W.E. (1991). Analysis of data from computer linked files. *Proceedings Interface 1991* (pp. 411-414). Fairfax Station, VA: Interface Foundation of North America.

Survival Analysis Methods for Medical Treatment Effectiveness Research

David C. Naftel, Ph.D.

Introduction

According to Relman (1988), the medical care system is entering the third revolution since the end of World War II. It will be characterized by assessment and accountability. Relman reports on a conference sponsored by the Health Care Financing Administration (HCFA) where representatives from Government, private insurers, major corporations, community agencies, and the medical profession demonstrated remarkable unanimity on the conclusion that the time had come for a new national health care initiative. They agreed that future decisions on funding and organization of health care must be based on increased knowledge of the variations in performance among institutions and practitioners and increased knowledge of the factors associated with these variations.

In its annual Medicare Hospital Mortality Information, HCFA (1991) carefully studies variations in outcome, specifically death, according to conditions and procedures for each hospital in the United States. They compute observed mortality and estimate risk-adjusted expected mortality. These estimates are based on a parametric survival model and are reported three times after admission: at 30, 90, and 180 days. Their analyses clearly demonstrate that variations do exist.

Assessment, accountability, and variations in performance are all components of the science of medical treatment effectiveness. Roper, Winkenwerder, Hackbarth, and Krakauer (1988) define medical effectiveness as "the appropriate use of an intervention in a given situation." The Agency for Health Care Policy and Research (AHCPR) (1990) poses the questions of medical treatment effectiveness in the following ways:

1. For a given condition in a given patient, is one type of operation more effective than another?
2. When is it appropriate to choose medical treatment rather than surgery?
3. When is it appropriate to watch and wait?

Closely related questions include:

1. When are the results from a specific institution (or physician) less than (or better than) the expected results?
2. When are results improving (or worsening) over time?

This paper presents a conceptual basis and the statistical estimation process (for an ideal quantitative representation) that can be, and have been, used to answer all of these questions. The underlying assumption is that the time-related probability of outcome (or equivalently the distribution of the time of outcome) is the most basic information of interest.

The ideas and methods presented in this paper are the result of 15 years of collaboration with Eugene H. Blackstone, Malcolm E. Turner, and John W. Kirklin. The methods are actually a synthesis of many known statistical methods and ideas. Three papers (Blackstone, Naftel, and Turner, 1986; Kirklin, Akins, Blackstone, and others, 1991; Blackstone, Naftel, and SAS Institute, Inc., 1988) contain details and examples of these methods. The examples in this paper are from data on treatment for ischemic heart disease.

Defining the Problem and the Ideal Answer

If more than one treatment exists for a specific patient with a given condition, then which treatment should be used? If a patient or group of patients has received a particular treatment, was it the appropriate treatment? If a physician or an institution has performed a number of certain procedures, have the resulting outcomes been consistent with certain estimated standards?

Ideally, the answers to these questions should be based on a time-related depiction of the probability of event (or the probability of freedom from the event) for each

Dr. Naftel is Associate Professor of Surgery, Department of Surgery, University of Alabama, Birmingham.

treatment. The depiction would be a function of patient-specific risk factors and treatment-specific risk factors. Such depictions allow complete time-related comparisons among treatments, including all of the usual comparison statistics such as 30-day survival, 1-year survival, 5-year survival, median survival, mean lifetime, hazard rates, and hazard ratios.

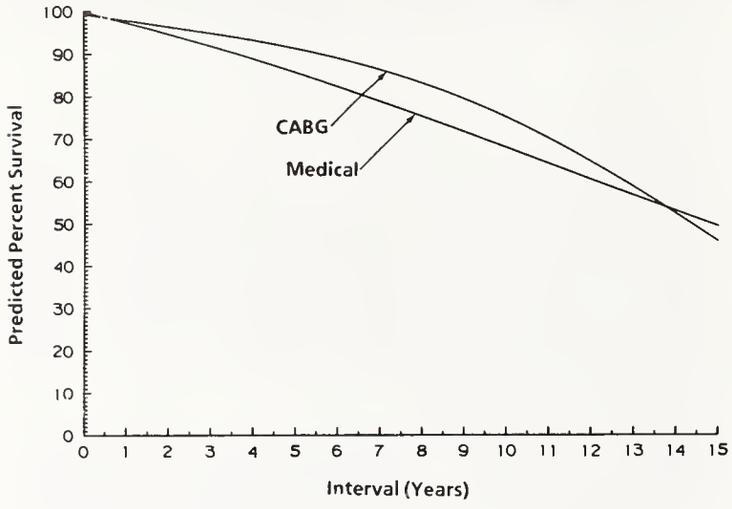
Consider a 55-year-old patient with ischemic heart disease who has specific values of risk factors such as three-vessel disease and moderate left ventricular dysfunction. The patient and his physician are considering either medical therapy or coronary artery bypass surgery (CABG). Figure 1 gives a hypothetical representation of the ideal information for this specific patient for the outcome death.

The medical line in Figure 1A represents the true (but unknown) probability of survival if medical therapy is chosen for this specific patient. Similarly, the CABG line represents the true probability of survival if CABG is chosen. Note that the 1-month survival is higher for medical therapy (99.8%) than for CABG (97.8%). At 5 years, however, the survival is higher for CABG (91%) than for medical therapy (86%). Then, at 14 years, the lines cross so that medical survival again becomes higher than surgical survival.

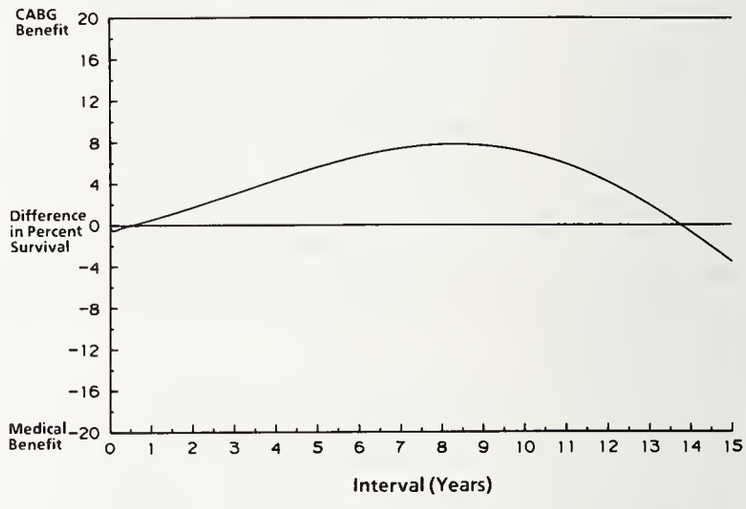
The shapes and magnitude of these two lines of probability and the resultant differences are functions of the patient-specific and treatment-specific risk factors. It is likely that the relationship of the lines may appear

Figure 1. Ideal depictions for time-related patient-specific treatment comparisons

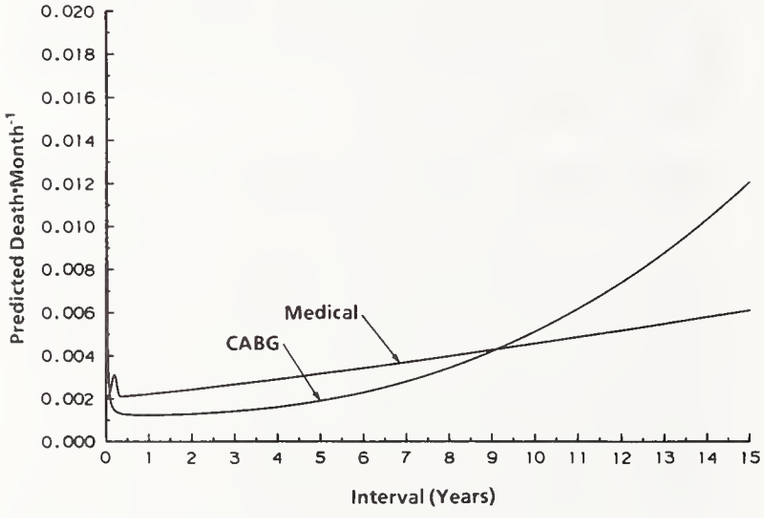
1A. Predicted probability of survival after intervention for coronary artery disease



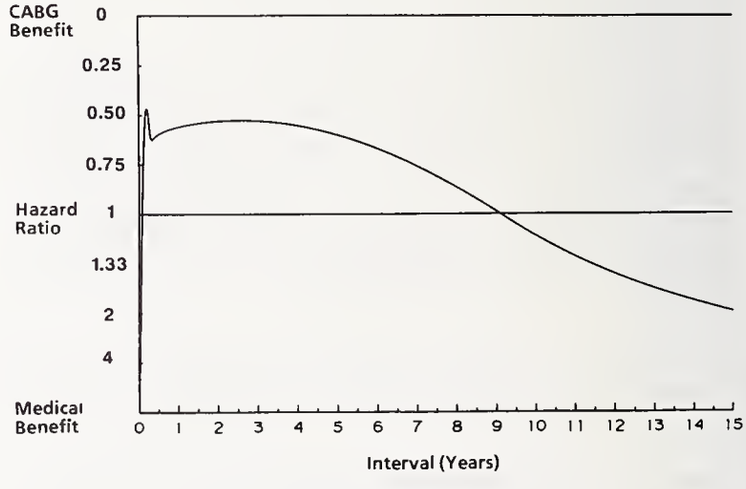
1B. Comparative benefit in terms of difference between predicted survival probabilities



1C. Predicted hazard functions (instantaneous death rate)



1D. Hazard ratio



Note: The patient is 55 years old with three-vessel disease and moderate left ventricular dysfunction.

entirely different for two patients with different risk factors.

Figure 1 contains additional information to visualize and quantify outcome (death) after each treatment. Figure 1B is the comparative benefit (difference in percent survival) between CABG and medical survival where points above the horizontal line of 0 represent survival in favor of CABG, and points below the line represent survival in favor of medical therapy. Additional insight is provided by the hazard function (instantaneous death rate) for each treatment (Figure 1C). Notice that the hazard function for medical therapy has a very short high early phase which merges to a constant rate. The CABG hazard function has a longer early phase that quickly merges to a constant rate and then begins to increase so that the two treatments have identical rates at 9 years. The ratio of the two hazard functions is given in Figure 1D. A ratio of 1 occurs when the two hazard functions are equal.

The information in Figure 1 would allow a truly informed decision. The decision would not be automatic but instead would depend on patient preferences concerning the relative importance of being alive at 1 month or 1 year or 10 years. There would be other considerations such as cost and time-related probabilities of other outcomes (for example, return of angina, subsequent intervention, return to work, and quality of life). It is entirely possible that two identical patients would make different decisions.

Figure 1 represents the true probabilities of death. In reality, these probabilities would never be known. They would be estimated from finite patient and time-related outcome information. The estimates would have associated measures of uncertainty that would translate to uncertainty associated with the selection of a treatment. Figure 2 is identical to Figure 1 except that it is derived from real data so that the solid lines are estimates of probability and the dashed lines are confidence limits for the estimates. The exact periods of time when a particular treatment is beneficial now have uncertainty associated with them, and concepts of statistical significance and confidence limits now have meaning and use.

The Process to Estimate the Answers

To produce ideal information depictions such as those presented in Figure 1, there must be clearly defined outcomes, a proper set of data including followup, statistical methods for estimating distribution shapes, statistical methods for identifying risk factors, and methods for graphically and quantitatively predicting results for a specific patient. The data and the methods must consider time until an outcome occurs.

Each step in this process is discussed in the following sections. The ultimate goal is to produce interpretable and informative results that will aid the decision process.

Outcome Events

One definition of outcome (Blumberg, 1986) is "those changes, either favorable or adverse, in the actual or potential health status of persons, groups, or communities that can be attributed to prior or concurrent care." With this definition, outcome events can be partitioned into three groups.

1. Discrete events that occur only once. Death is an outcome event that occurs only once. Other examples are a bit artificial but still useful. For example, the events "first return of angina," "first reoperation," or "first return to work" are all useful outcomes to analyze and would have great importance to the patient.

2. Repeating events. Many outcome events can occur multiple times. Reoperation after an initial CABG can conceivably occur several times. Episodes of rejection after heart transplant can occur many times throughout the life of the patient. Repeating events may have associated degrees of severity, such as degree of rejection.

3. Continuous events. Some outcome events are actually a continuous process over time. For example, cardiac output after cardiac surgery may change with time and indicate a worsening or an improving condition of the patient.

The vast majority of applications of statistical methods to outcome research involve the analysis of discrete events that occur only once, and the most common outcome is death. In fact, the area of statistical science that deals with outcomes is usually labeled survival analysis.

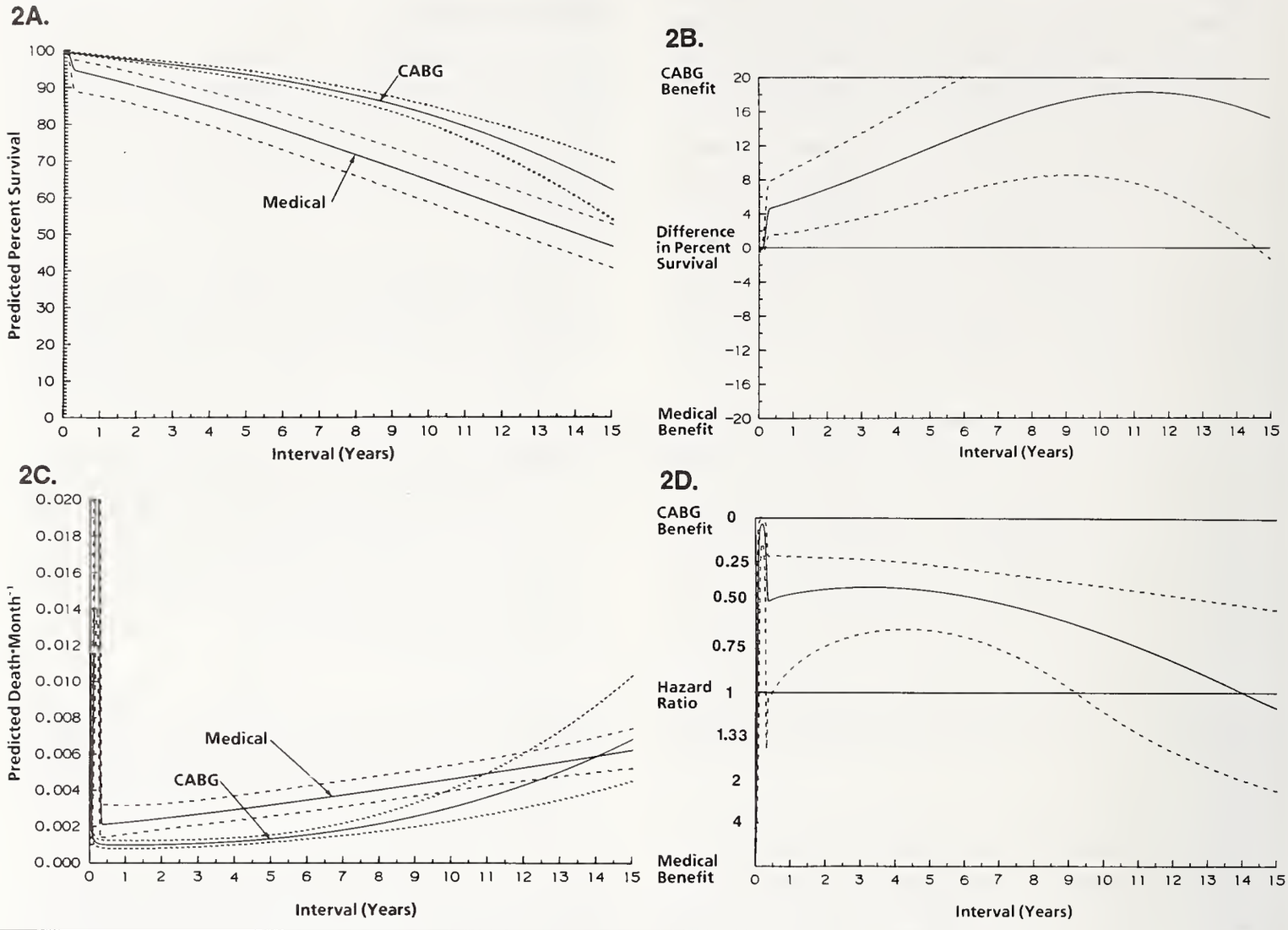
Data

The source of data and the method that generates the data are critical considerations of an analysis. Randomized trials are often cited as the ideal method to generate data for making treatment comparisons; however, even randomized trials have limitations. For example, many randomized trials enroll a narrow selection of patients such that inferences are limited in applicability.

This paper is not intended to discuss the many, many issues of producing data for an analysis. Rather the approach here is that if an investigator has data worth analyzing, then the data are worth an analysis that produces answers to the questions being asked. A full discussion of data issues is presented in the report to Congress by AHCPR (1991).

Nevertheless, it is simply unfair to expect any type of statistical analysis to purify, adjust, average, or in any

Figure 2. Nomograms depicting the predicted patient-specific comparative benefit of coronary artery bypass graft surgery and medical therapy for a 55-year-old male with three-vessel disease and moderate left ventricular dysfunction



Note: The form of the depictions is the same as in Figure 4.

way overcome data deficiencies. Therefore, the researcher must be aware of a few crucial considerations, especially if data from diverse sources are being combined and if time-related outcomes are being analyzed.

Definitions and standardization of variables. The outcome events must have precise, consistently applied definitions. This becomes especially important and difficult for outcomes that have an element of subjectivity such as cardiac death. Assigning a specific cause to a death is often difficult and many times, simply impossible for unobserved deaths and deaths with no autopsy.

Method of patient followup. The time of outcomes is known through patient followup. Important elements of patient followup are that it should be complete for all patients, it should identify precisely when each outcome occurred, and there should be no association between occurrence of an event and lack of followup (censoring) for a patient. A complete discussion of methods of followup is given by Blackstone (1988).

Era considerations. Comparing treatments is often complicated because one or both treatments may have been evolving over time. Also, as experience with a treatment grows, it is often applied to a wider group of patients so that the prevalence of risk factors may change with time. Statistical methods can deal with such realities to a limited extent by including date of treatment as a possible risk factor and by carefully examining treatment and data interactions.

Risk factor prevalence. The prevalence of the risk factors must be considered when inferences are drawn for an analysis. Studies with narrowly defined risk factors have limited usefulness because of the known dangers of extrapolating beyond the data.

Statistical Methods

The statistical methods for analyzing medical treatment effectiveness are most useful when time until the outcome is incorporated into the results. Such methods include:

1. Kapla-Meier (1958) product-limit estimates of survival — a time-related nonparametric estimate of the proportion surviving. These estimates and their standard errors provide the starting point for most survival analyses. They are especially useful in determining the form of the underlying survival distribution.

2. Nonparametric methods for comparing survival of two or more groups — survival after one of two treatments can be compared by one of many nonparametric statistical tests. Each of these tests produces a single p value that, when significant, essentially tells the researcher that there is a difference but gives little guidance about how the treatments differ. Most of these tests assume similar distributions except for a single parameter such as the median. These tests are not useful when the distributions are quite different in shape and the nature of this difference is the pertinent information. These tests do not adjust for risk factors. (See Elandt-Johnson and Johnson, 1980, for an overview.)

3. Subsetting of data for actuarial survival estimates for patient subgroups of interest — a common approach to recognizing the effect of known or hypothesized risk factors is to subset the data set into risk groups (Killip, Pasamani, Davis, and the CASS Principal Investigators and their Associates, 1985). For example, patients who have undergone CABG might be subsetted into three groups based on the number of coronary systems that were diseased at the time of surgery. Each of these three groups may be further subsetted according to age decades. The number of such subsets quickly becomes very large as the number of stratification variables increases. To be practical, this approach requires a very large number of patients. It is also difficult to produce a clear unified picture of the effect of risk factors from this approach.

4. Cox proportional hazard regression analysis — Cox had a major impact on survival analysis with his introduction of proportional hazard regression analysis (Cox, 1972). His method estimates the proportional effect of risk factors on the underlying unknown hazard function.

5. Parametric survival analysis — historically, parametric survival analysis has involved relatively few distributions and few parameters. Concomitant information is usually added to the model by making the scale parameter a linear or log-linear function of concomitant variables. In an early example of this approach, Feigel and Zelen (1965) used a constant hazard model to analyze survival data. A constant hazard corresponds to a negative exponential density. The parameter for the constant hazard was made to be a linear function of a risk factor. Other researchers who have investigated models that are generalizations of the negative exponential include Weibull, Gamma, Rayleigh, and Makeham. Cox and Oakes (1984) give the equations for many of these distributions.

HCFA (1991) used the modified Makeham distribution to estimate time-related survival according to hospital and diagnosis. That analysis let both a scale and a dispersion parameter be functions of concomitant information.

Multiphase Parametric Survival Analysis

A system of parametric survival analysis which includes most of the standard parametric models was developed by Blackstone, Naftel, and Turner (1986). This system was developed to provide a very general, mathematically tractable, statistically robust system of parametric survival distribution equations that would be applicable to a wide range of possible survival distribution shapes. The model belongs to the class of parametric survival models known as mixture distribution models. The distribution of events can be considered as a mixture of as many as three distribution models referred to in this paper as "phases." Although all phases are defined over all time, each phase predominates in only one timeframe and therefore can be expressed in rather simple form with few estimated parameters. Each phase consists of a parametric shaping distribution function scaled by a parametric scaling function of concomitant information. Each shaping distribution function is a general one that can be thought of as giving rise to a set of nested hierarchical submodels, with the simplest one that adequately fits the data being the one sought. Thus, many of the parametric models currently in use are either submodels of the system or can be approximated closely with one.

The mathematical form of the model has been developed with considerable care to give it attractive statistical and mathematical properties for maximum-likelihood estimation. Although the detail of the full system of equations designed to accommodate a wide range of possible distribution shapes is of necessity complex, the

specific distribution model developed for a set of data is usually a simple one with few shaping parameters.

The wide diversity of the distributions within this family of distributions is apparent by some of the special cases which include the exponential, Rayleigh, generalized Weibull, Weibull, Gompertz, Makeham-Gompertz, and modified Makeham. Thus, the hazard function can be constant, decreasing, increasing, increasing to a maximum and then decreasing or any combination of these features.

The parametric system is implemented by two user-contributed procedures, PROC HAZARD and PROC HAZPRED, in the SAS software system (Blackstone, Naftel, and SAS Institute, Inc., 1988). There are three main functions of the system:

1. Modeling — the system can be used to explore and implement any one of many parametric models. The most general model consists of three additive hazard phases that can model early, constant, and late behavior of the distribution. All estimation is through maximum likelihood.

2. Risk factors — each of the three hazard phases is scaled by a log-linear function of risk factors. Risk factors are examined by an extremely detailed and controllable stepwise regression package.

3. Predictions — once a model and accompanying risk factors are determined, then time-related, patient-specific estimates of probability of death or survival and hazard and cumulative hazard functions can be calculated and plotted. Confidence limits for all predictions are given. Such predictions and confidence limits are extremely valuable for graphic assessment of the effect of risk factors or for comparing alternative treatments.

Identification of Shapes (Distributions)

When two treatments that differ greatly in their impact on a patient are compared (for example, CABG and medical therapy), there is no reason to expect similar distributions of outcome across time. In fact, it is well known that the trauma of surgery leads to a high early death rate that is quite different from the early death rate of medical therapy. Other parts of the distribution usually differ in shape also (see Figure 1).

These different shapes of the distributions greatly complicate statistical comparisons. Most statistical methods for comparing two groups—that is, treatments—make the fundamental assumption that the treatments have identical distributions except for a single parameter of interest, such as a location parameter. For many applications, this assumption is reasonable. If the two treatments have entirely different distributions, however, then the question is no longer "are the treatments different with respect to outcome?" but rather "at what points in time are the treatments similar and at what points in time are they different?"

The first step in comparing treatment distributions, or equivalently, the probability of outcome across time, is to select a statistical distribution that is appropriate for each treatment. The flexible family of distributions presented in the last section is useful for this purpose. The specific parsimonious family member can be chosen through maximum likelihood estimation, likelihood ratio tests, and graphic comparisons of nonparametric and parametric survival estimates.

Estimation of the Effect of Risk Factors

Reduction of a list of concomitant variables to a useful set of risk factors is an intense task that requires both medical and statistical expertise. The parametric system described in this paper explicitly recognizes that a risk factor may exist in one phase of the time-related hazard but not in another. Other than that important detail, a stepwise regression in the parametric setting is similar to regression in other settings. A large body of literature is available on the strategies, benefits, and precautions of multiple stepwise regression in both the linear regression setting (Draper and Smith, 1966) and the survival analysis setting (Elandt-Johnson and Johnson, 1980). This literature is applicable to regression of risk factors in parametric survival analysis and serves as a reminder that the analysis cannot be totally automated.

The general approach for estimating risk factors for groups of patients with two different treatments is:

1. Determine the distribution for each treatment group.

2. If the distributions are similar except for scale parameters, then investigate risk factors, including treatment variables, in the entire group. Be aware of interactions of risk factors with treatment.

3. If the distributions have different shapes, then investigate risk factors separately for the two treatment groups.

Elandt-Johnson and Johnson (1980) give a detailed description of this approach, as do Blackstone, Naftel, and SAS Institute, Inc. (1988).

Examples

The examples in this section illustrate the steps involved in applying the multiphase parametric system and present the results of the analysis.

Example 1. Time-related probability of death after CABG. A study of 5,880 patients undergoing CABG was analyzed and the basic results presented (Kirklin,

Akins, Blackstone, and others, 1991; Sergeant, Lesaffre, Flameng, and Suy, 1990). Figure 3A contains both the nonparametric and the parametric estimates of survival for the entire group of patients. Figure 3B contains the corresponding parametric hazard function. The hazard function, as estimated from the data, has three phases. The high early phase is essentially over by 3 months. The constant phase is joined by a rising late phase whose influence becomes apparent at about 6 years.

The next step was the identification of risk factors associated with each phase. After many potentially important variables were examined in a stepwise regression fashion for all three phases simultaneously, three lists of risk factors were produced (Table). The effect of each risk factor can best be described by nomograms. For example, diabetes is a risk factor whose effect is in the early and late phases. Figure 4 contains solutions of the multivariate equation. All risk factors except diabetes are set to values for a typical patient. Notice that the early effect of diabetes is small and the late effect is quite large, so that 15-year survival after CABG is predicted to be 36 percent for a diabetic patient and 79 percent for a nondiabetic patient.

The effect of a particular aspect of the operation can also be examined. Nonuse of the internal mammary artery (IMA) (usually for revascularizing the left anterior descending artery) was found to be a risk factor in the late phase. The effect of using or not using the IMA is shown in Figure 5.

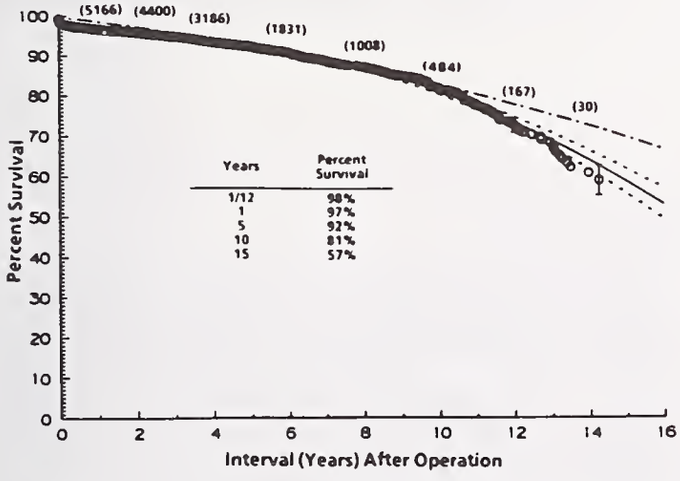
Example 2. Time-related probability of death after initial medical therapy. A study of 354 patients with coronary artery disease who were randomly assigned to initial medical therapy was analyzed (Kirklin, Akins, Blackstone, and others, 1991; Veterans Administration Coronary Artery Bypass Surgery Cooperative Study Group, 1984). The parametric analysis supported the existence of three phases of hazard: early, constant, and late. Left main disease was among the risk factors identified by stepwise parametric regression. Its effect was significant in the early and constant phases. Figure 6 illustrates the effect of left main coronary artery disease for a typical patient. Notice that left main disease is a risk factor in the early and constant phases of hazard.

Example 3. Comparing treatments for coronary artery disease. Consider a 56-year-old male patient with three diseased coronary systems and left main disease and no history of myocardial infarctions. The patient and physician are now considering surgery or medical therapy. Figure 7 contains the time-related predictions of probability of death and predictions of hazard (instantaneous death rate) for this specific patient. For this patient, the decision is clear. Surgery offers the best probability of survival across time.

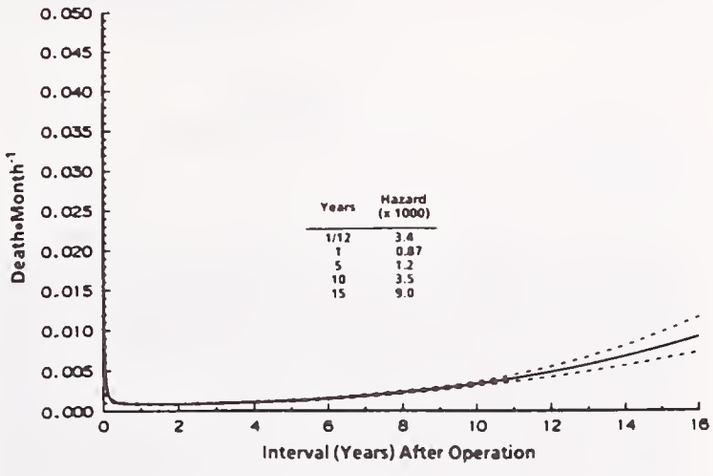
If this same patient did not have left main disease (Figure 8), then the decision in favor of surgery is still clear, but the early benefit of surgery is less, so that postponement of surgery might be acceptable.

Figure 3. Estimated probability of survival and estimated hazard function (instantaneous risk of death) in 5,880 patients after coronary artery bypass graft surgery

3A. Time-related survival probability after coronary artery bypass graft surgery



3B. Hazard function with three phases of hazard



Note: The circles are nonparametric estimates, and the solid line is a nomogram of the parametric estimated probability of survival. The dashed lines enclose the 70 percent confidence intervals. The dash-dot-dash line is the survivorship of an age-gender-race match general population.

Source: Reproduced with permission from Kirklin, Akins, Blackstone, and others (1991).

Table. Incremental risk factors for death after primary isolated coronary artery bypass graft surgery

Incremental risk factors for death (deaths = 545)		Hazard phase		
		Early	Constant	Late
Older	Age	•		•
Younger	Age			•
	Distribution of coronary artery disease			
Greater	Number of diseased vessels			•
	Left main disease > 90 percent	•		
	Left ventricular function			
Lower	Ejection fraction	•	•	•
	Left ventricular hypertrophy	•		
	Ischemic mitral incompetence		•	
	Angina			
	No anginal symptoms		•	
Greater	Severity of angina	•		
	Unstable angina	•	•	
	Nonexertional angina	•		
	Myocardial infarction			
	Hemodynamic instability	•		
	Cardiogenic shock	•		
	Atherosclerotic disease			
	Peripheral vascular disease	•	•	•
More	Small coronary arteries	•		•
	Coexisting conditions			
	Insulin-treated diabetes	•		•
Lower	1-second expiratory rate (percent of normal)	•		•
	Hyperlipidemia			
Lower	HDL levels		•	
Higher	Triglyceride levels			•
	Surgical factors			
	Non-use of IMA			•
More	Number of endarterectomies	•		
More	Use of sequential grafts	•		
	Incomplete revascularization		•	
	Environmental factors			
	Surgeon B	•		
Earlier	Date of operation			

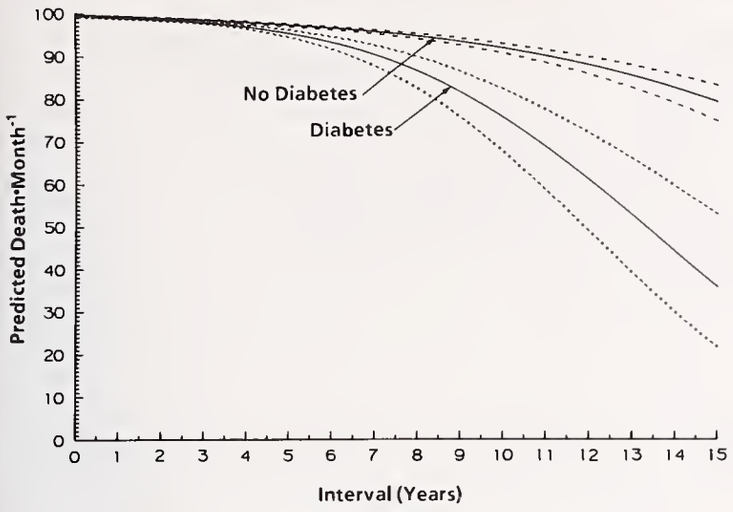
The next patient is a 55-year-old male with three-vessel disease and moderate left ventricular dysfunction. Figure 2 contains the predictions for this patient. Notice that the survival probabilities cross twice, once at 9 months and again at 13 years. Patient preferences would be an important part of the decision. To maximize the probability of living for 6 months, the patient should choose medical therapy. If 5-year survival is of the greatest importance, then the patient should choose surgery. These decisions should be re-evaluated as symptoms progress and treatments evolve.

Discussion

Learning about treatments. Estimation of the probability of outcome across time is important in learning about the effect of a treatment. A treatment may have a high early risk and then a long period of low risk. Another treatment may have low initial risk but a quickly rising late risk. Incorporation of risk factors into the probability of outcome across time allows a better understanding of the effect of risk factors. A risk factor may be important in only one phase of hazard, and in fact, a low

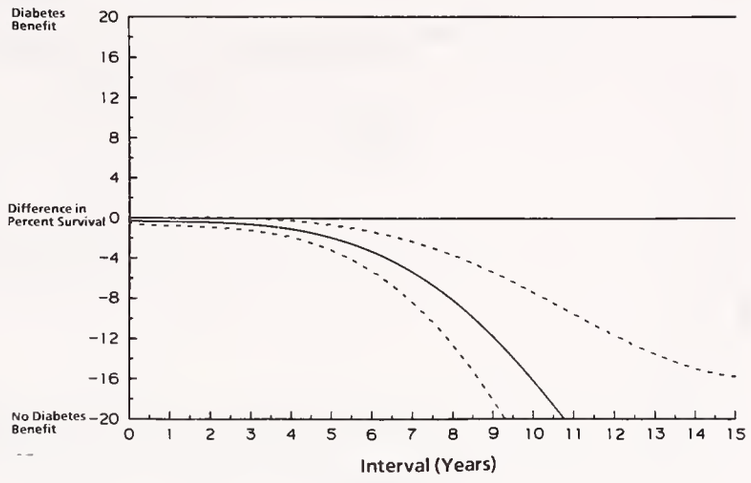
Figure 4. Nomograms depicting the predicted effect of diabetes after coronary artery bypass graft surgery

4A. Predicted probability of survival



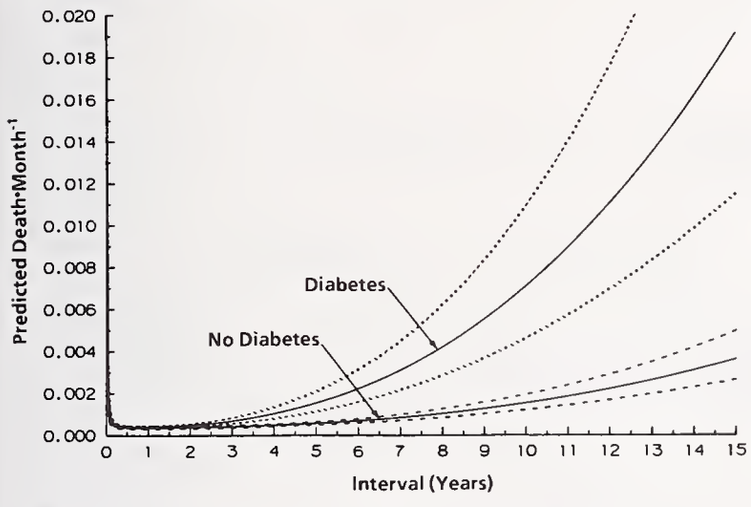
Note: Dashed lines enclose the 70 percent confidence limits.

4B. Comparative benefit in terms of survival differences



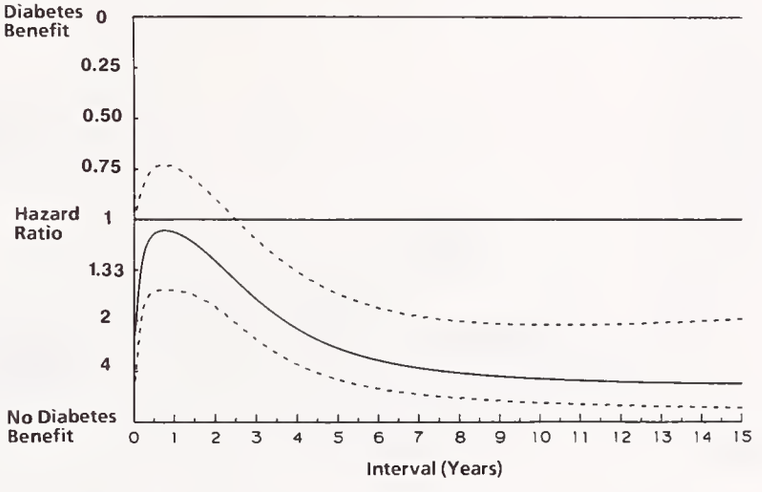
Note: Dashed lines enclose the 90 percent confidence limits to facilitate comparisons with 0.

4C. Predicted hazard function



Note: Dashed lines enclose the 70 percent confidence limits.

4D. Hazard ratio



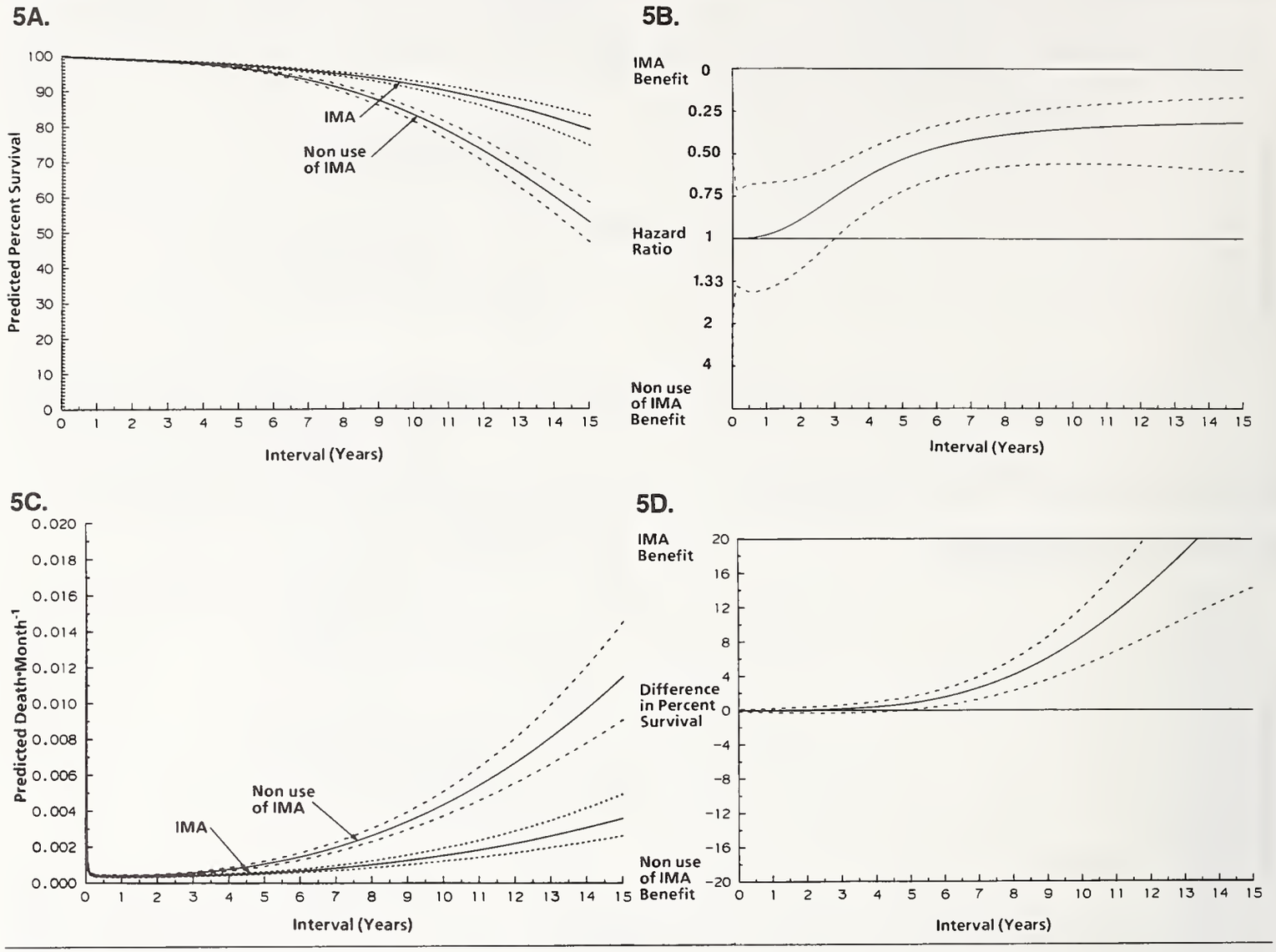
Note: Dashed lines enclose the 90 percent confidence limits to facilitate comparisons with 1.

value of the risk factor may essentially negate the increased risk of the phase. This kind of detailed knowledge is the first step in mediating the effect of controllable risk factors.

Patient-specific treatment comparison. It is possible that for a given disease, treatment A is better than treatment B for all patient circumstances, so that treat-ment B should be totally abandoned in favor of treatment A. Or there may be definable subgroups of patients where a clear treatment winner exists. For example, the Coronary Artery Surgery Study (Killip, Pasamani, Davis, and the CASS Principal Investigators and their Associates, 1985) which was designed to compare surgical and medical treatment for coronary artery disease did not randomize some patients because "...it was already apparent by 1973 that bypass surgery markedly improved symptoms in severely symptomatic patients."

For many diseases, however, and many patients, a decision among treatments is not obvious. A randomized

Figure 5. Nomograms depicting the predicted effect of using the internal mammary artery for revascularization



Note: The form of the depictions is the same as in Figure 4.

trial that produces a single p value is often of little help when treatment for a specific patient needs to be recommended.

According to Elandt-Johnson and Johnson (1980), "...survivorship is a continuous process, and comparisons based on single statistics are not really appropriate." Time-related depictions of the probability of outcome adjusted for the traits of a specific patient are most helpful. Such depictions allow time-related comparisons between treatments including the magnitude of the differences and the uncertainty associated with the differences. Proper application of parametric survival analysis provides these depictions.

These parametric methods are not intended to replace clinical judgment. Rather, they are intended to provide information, based on past studies, in a manner that is patient specific and usefully depicted. These depictions can also provide the basis for truly informed consent.

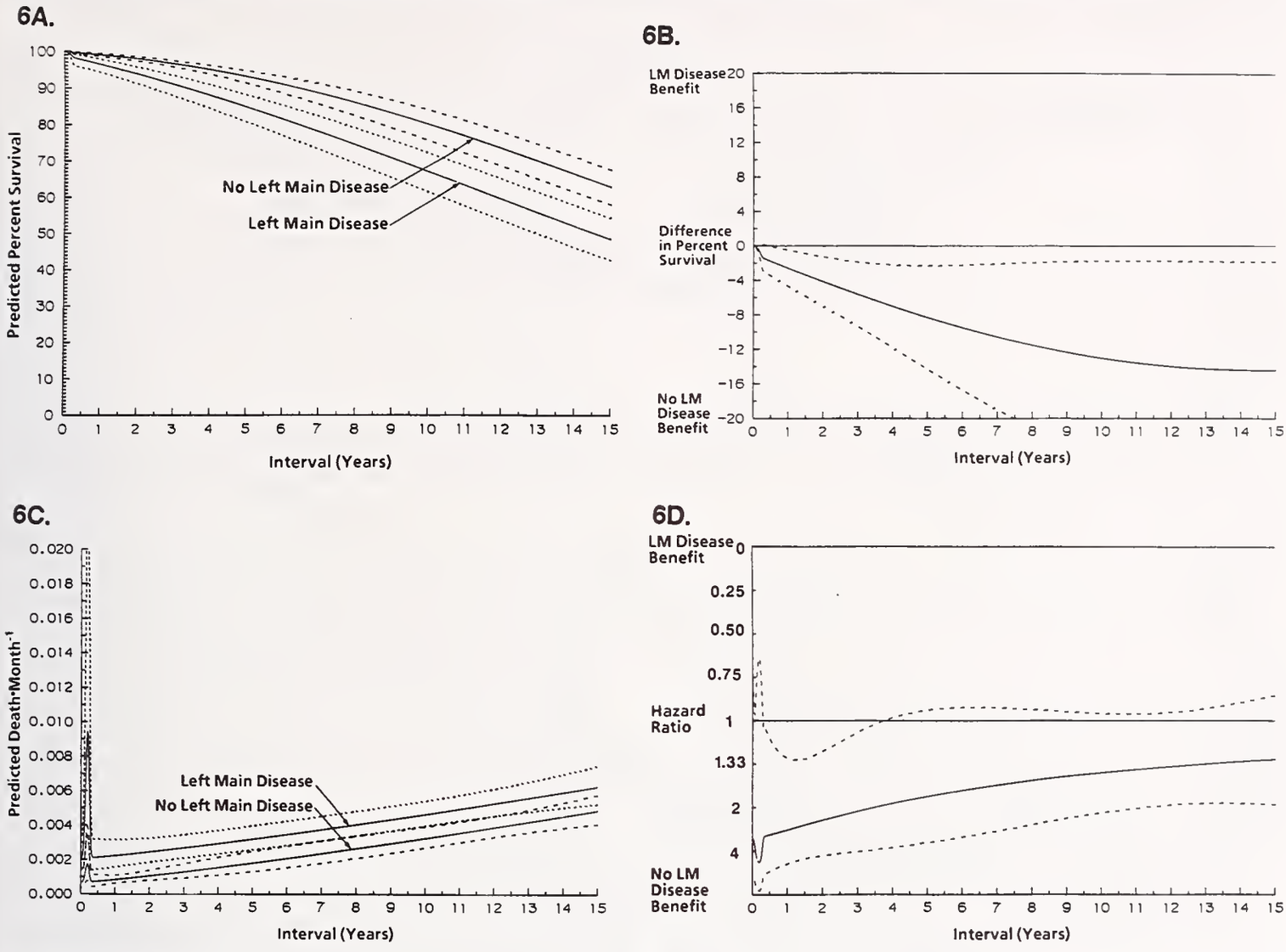
Appropriateness. Parametric survival analysis is ideally suited for assessing appropriateness. The actual observed number of outcomes in a group of patients receiving a specific treatment can be compared with the expected number of outcomes had the patients received another treatment. The expected number of events is a function of each patient's risk factors and his length of followup. This methodology works similarly in comparing the results of a particular surgeon or a particular hospital with results from another surgeon or hospital.

Conclusion

Parametric survival analysis including risk factor estimation is a useful statistical technique that can accomplish the following:

1. Estimate the pattern (distribution) of the probability of outcome across time and estimate the phases of risk;

Figure 6. Nomograms depicting the predicted effect of left main (LM) disease after medical therapy

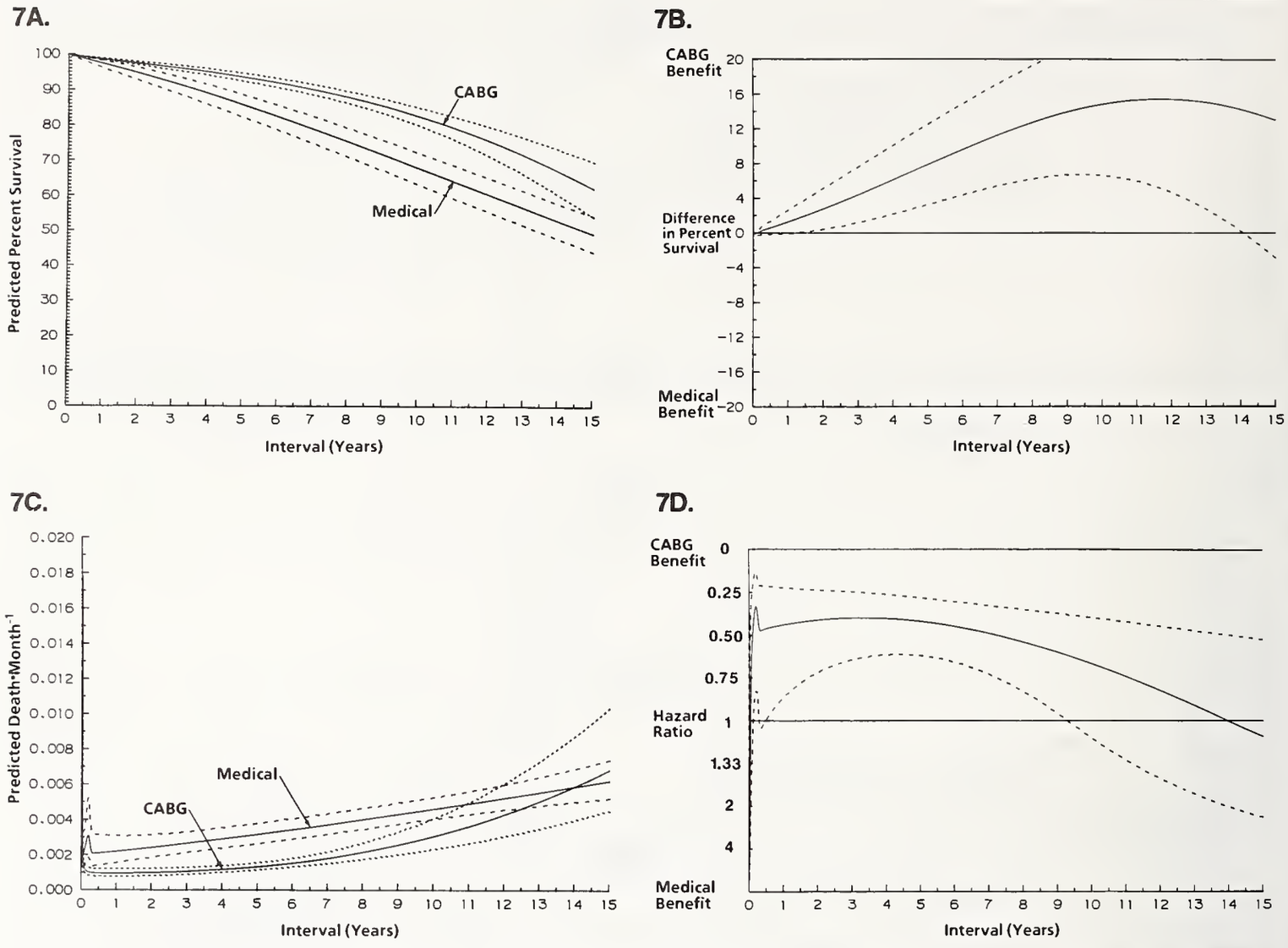


Note: The form of the depictions is the same as in Figure 4.

2. Identify and graphically depict the effect of risk factors on probability of outcome;
3. Allow risk-adjusted comparisons of treatments or comparisons of aspects of the same treatment; compare other factors, for example, hospitals, surgeons, and eras;
4. Predict the future time-related probability of death for a specific patient; and
5. Assess appropriateness of past treatment decisions based on time-related observed versus expected outcomes.

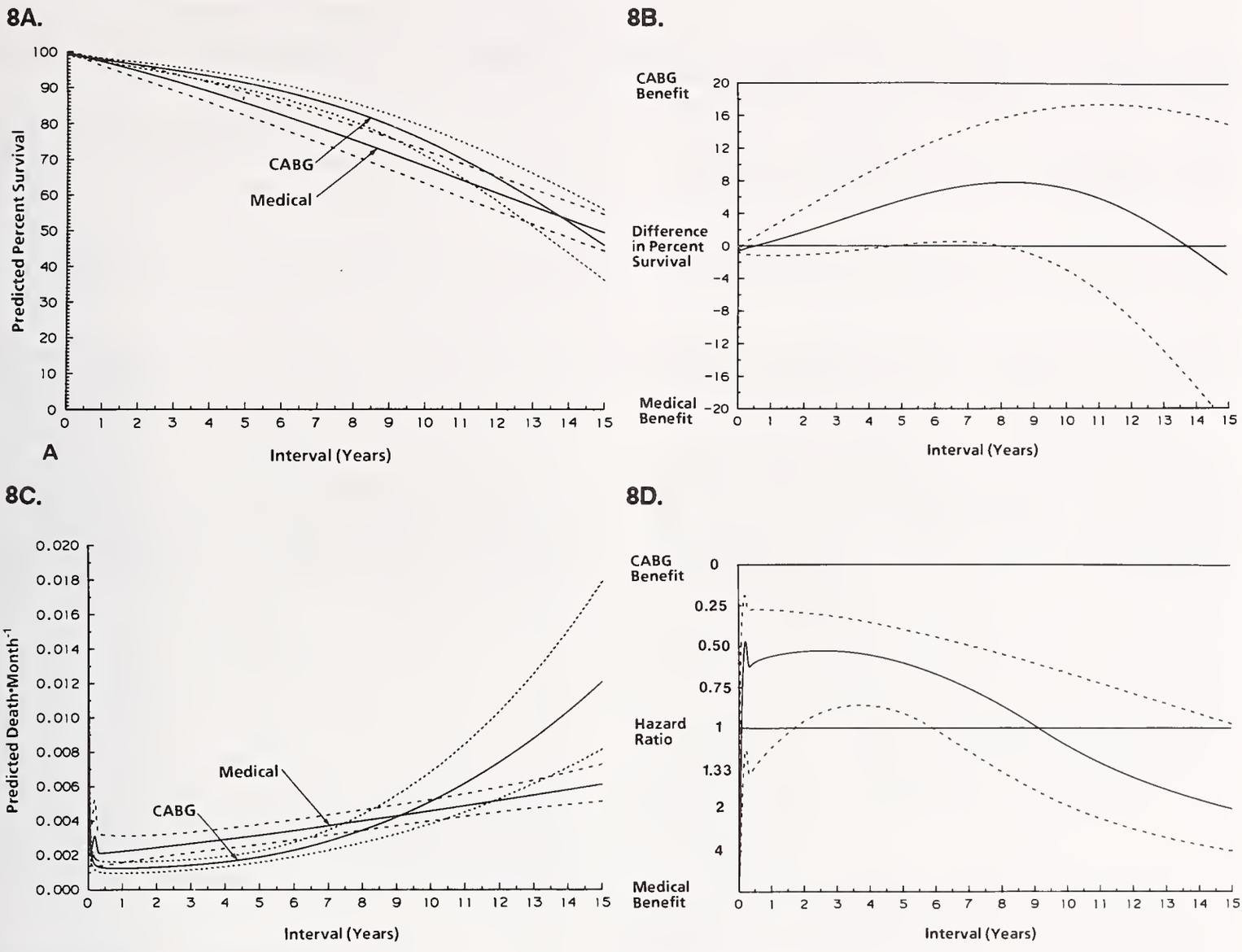
Only properly applied parametric survival analysis can accomplish all of these tasks in a clear and graphically useful manner. The parametric system presented in this paper is especially well adapted to cover a wide variety of time-related outcome data.

Figure 7. Nomograms depicting the predicted patient-specific comparative benefit of coronary artery bypass graft surgery and medical therapy for a 56-year-old male patient with three-vessel disease and left main disease



Note: Other risk factors are set at typical values. The form of the depictions is the same as in Figure 4.

Figure 8. Nomograms depicting the predicted patient-specific comparative benefit of coronary artery bypass graft surgery and medical therapy for a 56-year-old male patient with three-vessel disease and no left main disease



Note: Other risk factors are set at typical values. The form of the depictions is the same as in Figure 4.

References

Agency for Health Care Policy and Research. (1990). *Patient outcomes, research teams, and the Agency for Health Care Policy and Research* (Program Note). Rockville, MD: Author.

Agency for Health Care Policy and Research. (1991). *Report to Congress: The feasibility of linking research-related data bases to Federal and non-Federal medical administrative data bases.* Rockville, MD: Author.

Blackstone, E.H. (1988). Analysis of death (survival analysis) and other time-related events. In Macartney, F.J. (Ed), *Congenital heart disease* (pp. 55-101). London: MTP Press, Ltd.

Blackstone, E.H., Naftel, D.C., and Turner, M.E. Jr. (1986). The decomposition of time-varying hazard into phases, each incorporating a separate stream of concomitant information. *Journal of the American Statistical Association*, 81, 615-624.

Blackstone, E.H., Naftel, D.C., and SAS Institute, Inc. (1988). *Changes and enhancements to the SAS system, Release 5.18, under OS and CMS* (SAS Technical Report P-175). Cary, NC: SAS Institute, Inc.

Blumberg, M.S. (1986). Risk adjusting health care outcomes: A methodologic review. *Medical Care Review*, 351-393.

Cox, D.R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society*, 34, 187-220.

Cox, D.R. and Oakes, D. (1984). *Analysis of survival data.* London: Chapman and Hall Ltd.

Draper, N.R. and Smith, H. (1966). *Applied regression analysis* (2nd ed.). New York: John Wiley & Sons.

Elandt-Johnson, R.C. and Johnson, N.L. (1980). *Survival models and data analysis.* New York: John Wiley & Sons.

Feigel, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrika*, 21, 826-838.

Health Care Financing Administration. (1991). *Medicare hospital mortality information, 1987, 1988, 1989* (vol. 25, Technical supplement). Washington: Author.

Kaplan, E.L. and Meier, P. (1958). Non-parametric estimation from incomplete observations. *American Statistical Association,* 53, 457-481.

Killip, T., Pasamani, E., Davis, K., and the CASS Principal Investigators and their Associates. (1985). Coronary Artery Surgery Study (CASS): A randomized trial of coronary bypass surgery. Eight years follow-up survival in patients with reduced ejection fraction. *Circulation*, 72(Suppl. V), V-102 to V-109.

Kirklin, J.W., Akins, C.W., Blackston, E.H., and others. (1991). Guidelines and indications for coronary artery bypass graft surgery. A report of the American College of Cardiology/American Heart Association Task Force on Assessment of Diagnostic and Therapeutic Cardiovascular Procedures (Subcommittee on Coronary Artery Bypass Graft Surgery). *Journal of the American College of Cardiology*, 17, 543-589.

Relman, A.S. (1988). Assessment and accountability: The third revolution in medical care. *New England Journal of Medicine*, 319, 1220-1222.

Roper, W.L., Winkenwerder, W., Hackbarth, G. M., and Krakauer, H. (1988). Effectiveness in health care: An initiative to evaluate and improve medical practice. *New England Journal of Medicine*, 319, 1197-1202.

Sergeant, P., Lesaffre, E., Flameng, W., and Suy, R. (1990). Internal mammary artery: Methods of use and their effect on survival. *European Journal of Cardiothoracic Surgery*, 4, 72-78.

The Veterans Administration Coronary Artery Bypass Surgery Cooperative Study Group. (1984). Eleven-year survival in the Veterans Administration randomized trial of coronary bypass surgery for stable angina. *New England Journal of Medicine*, 311, 1333-1339.

Problems and Potentials in Modeling Survival

L. Richard Smith, Ph.D.
Frank E. Harrell, Jr., Ph.D.
Lawrence H. Muhlbaier, Ph.D.

Introduction

The three general goals of survival analysis are to: compare survival curves among two or more therapies, elucidate the factors that influence survival and estimate their effects, and estimate the future survival for an individual or for a population. In medical effectiveness research, these three components are inextricably linked. Naftel (in this volume) presents a compelling framework for accomplishing these goals through the use of a powerful parametric survival model. Although his discussion centers on a particular parametric model, the principles propounded are relevant to all forms of survival modeling.

This paper will discuss treatment comparison, covariate selection, and prediction in the context of any survival model, in addition to the strengths and weaknesses of the multiphase model and some implications of survival analysis. A brief appendix describes the multiphase model.

Comparing Treatment Effects

Comparing two treatments is relatively easy if both treatments are from the same underlying hazard function family. Then the underlying hazard function can be modified by covariates in a proportional hazards manner or through direct modification of one or more of the parameters of the hazard function (see appendix). If each treatment has a different form of hazard function, there is no generally accepted method. If a different hazard function is assumed for each treatment, one can only easily test for factors affecting treatment differences by stratifying on treatment and then testing for covariate-by-treatment interactions (Thall and Lachin, 1986; also see the appendix to Califf, Harrell, Lee, and others, 1989, for a description).

Naftel (in this volume) uses an alternative strategy. The model is fit to each treatment's data. If the underlying hazard functions are different, that is prima facie evidence that the treatments have different time-varying outcomes. The question then becomes, "At what times do the survival curves differ?" The predicted survival and 70-percent confidence limits are produced for each treatment for an individual patient based on the patient's particular covariate values. The two survival curves are then compared across any time points of interest. Where the curves are distinctly different—that is, where the confidence limits are widely separated or, equivalently, where the 95-percent confidence interval for the survival difference does not include zero—the treatments are declared different.

A classical statistician would find this approach troublesome. Examining each time point for significance may yield spurious conclusions due to the multiplicity of tests. The counterargument is that we have already found the treatments to be different; we are now only interested in where they are "practically" different. The multiplicity-of-tests problem is mitigated by several factors. First, the model tends to produce "smooth" estimates of the prediction curves, and thus differences between adjacent time points tend to be similar; there are no erratic adjacent differences. A further mitigating factor is the knowledge of the disease and its treatment and the understanding of modeling process the physician and statistician bring to the problem at hand. The knowledgeable investigator is unlikely to make too much of an isolated difference.

Covariates

An area not discussed at length in Naftel's paper (in this volume) but essential in any survival analysis is the difficult problem of choosing appropriate covariates. Four general areas relate to covariates: selection and re-

Dr. Smith is Assistant Professor, Dr. Harrell is Associate Professor, and Dr. Muhlbaier is Research Assistant Professor; all are with the Division of Biometry and Medical Informatics, Department of Community and Family Medicine, Duke University Medical Center.

liability, composite covariates or indices, the functional form of the covariate, and the number of covariates.

Selection and reliability. In most situations, many covariates influence outcome in addition to treatment. One common method of determining relevant covariates is through the use of a stepwise regression method. This method is systematic and objective, but as in all other aspects of the modeling process, there are dangers. The process of choosing the final set of covariates requires many tests of significance—the multiple comparison problem—that may lead to a large Type I error.

To understand the stability of the variable selection process, Efron and Gong (1983) proposed "bootstrap" methods, in which the data set is sampled with replacement m times, where m is some suitably large number. As an example, one of us (Harrell) ran a simulation for selection of covariates relating to time until death on a population of critically ill patients (N = 984, deaths = 574). There were 18 candidate variables (denoted here only by the numbers 1-18). Thirty resampling runs were done. The variables selected are shown in Table 1. The number of variables selected ranged from 3 to 10 (Figure 1). Some variables were selected only once, but no variable was selected every time. The purpose of this simulation is not to suggest removing stepwise regression from the repertoire of modeling techniques, but to urge caution in uncritically accepting the results.

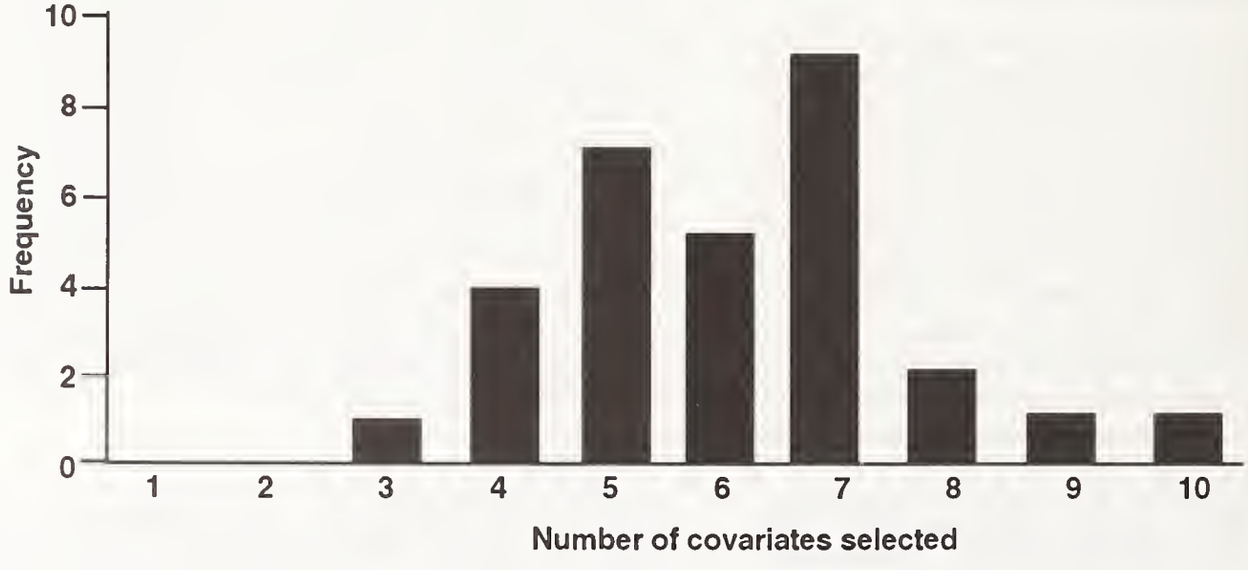
Several methods can be used to examine the reliability of the model and the reliability of the covariate estimates. For the overall model, one simple method is to plot the average model-predicted survival estimate against the corresponding Kaplan-Meier survival estimate for each decile of predicted survival. Estimates of survival can be analyzed similarly for subsets defined by covariates (for example, diabetics and nondiabetics), thus giving an estimate of how well the model predicts for a particular covariate. Note that in this latter case, the covariate defining the subset does not have to be a covariate included in the model. Sex may not affect survival, but a good model would predict well for both the male and female subsets. Because models developed on a particular data set generally estimate well within the same data set, however, additional methods should be employed.

One of these methods is data splitting. In its simplest form, the strategy is to divide the data set randomly into two subsets: a training data set and a test data set. The model developed on the training data set is then applied to the test set. Comparing the model predictions in the test data set with Kaplan-Meier estimates of survival will yield an estimate of model reliability.[1] Data splitting results in discarding a large fraction of the data for both model fitting and estimation of model reliability and, hence, is inefficient.

Data splitting can be improved by k-group cross-validation. This is done by subdividing the data set into k mutually exclusive and exhaustive subsets. The k separate models are developed using stepwise variable selection on the (k-1)/k fraction of the data, and the model is validated on the remaining 1/k fraction of the data. Es-

[1]This is also being done in a larger framework. There is considerable effort through the Chronic Ischemic Heart Disease Patient Outcomes Research Team (Pryor, D.B., principal investigator, AHCPR grant no. HS06353-01, Outcome assessment program in ischemic heart disease, 1989) to test models developed at one institution on data sets collected at other institutions.

Figure 1. Number of covariates selected in resampling simulation



Note: N = 984, deaths = 574

Source: Duke University Heart Center.

Table 1. Variables selected in a resampling simulation

	Variable																		
Run	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	n
1	•			•		•													3
2		•	•	•					•			•	•	•	•				8
3	•		•	•					•			•				•	•		7
4		•	•	•					•			•			•	•			7
5				•			•		•	•		•							5
6		•	•					•	•		•	•			•	•			8
7		•	•	•					•			•					•		6
8	•		•	•			•		•			•			•				7
9			•	•			•		•	•		•					•		7
10		•		•					•			•							4
11		•	•	•					•	•	•	•							7
12	•	•		•				•	•			•				•			7
13	•	•	•	•				•	•	•	•	•				•			10
14		•	•	•	•		•		•			•							7
15	•	•	•	•								•	•						6
16			•	•					•			•							4
17		•							•	•	•	•							5
18		•		•	•				•			•							5
19		•		•					•	•		•							5
20	•	•		•					•			•							5
21	•	•	•	•					•		•	•	•					•	9
22	•	•		•														•	4
23		•		•					•			•							4
24				•			•			•		•	•			•		•	7
25	•	•	•	•					•	•		•							7
26		•	•						•			•	•						5
27	•		•	•					•			•			•				6
28		•		•			•		•	•		•							6
29				•	•					•		•						•	5
30		•		•					•	•	•	•							6
N	11	21	16	27	3	1	6	3	25	11	6	28	5	1	5	6	3	4	

Run = Simulation run number
n = Number of variables selected for the run
N = Number of times a specific variable was selected

timates of fit (for example, R^2, difference between predicted and observed survival) are then averaged over the k validation runs. If the model validates, it is then fit to the entire data set for final parameter estimates. If the model does not validate, careful examination of the underlying hazard and candidate covariates should be undertaken.

Another method uses the bootstrapping technique (Efron, 1982). Here, the same set of data can be used to both fit and validate the model. Bootstrapping allows one to estimate and correct for the tendency of the predicted estimate to correlate better with Kaplan-Meier survival estimates than would be expected if the validation were done on separate data sets.

Indexes. A second consideration in variable selection is that several variables may jointly describe a particular aspect of a patient's condition, such as heart function or anginal pain. Because such an aspect of a patient's condition may manifest itself in different ways in different patients, the inclusion of all related variables in the model may dilute the overall effect of any individual component. Combining variables that describe the same condition into a composite index can yield more powerful results than using the individual variables. Harrell, Lee, Califf, and others (1984) gave an example in which 30 variables were reduced to 10 indexes. As an example, anginal symptoms in Table 2 form a composite based on three separate measures of anginal pain: severity, type, and frequency (Califf, Mark, Harrell, and others, 1988).

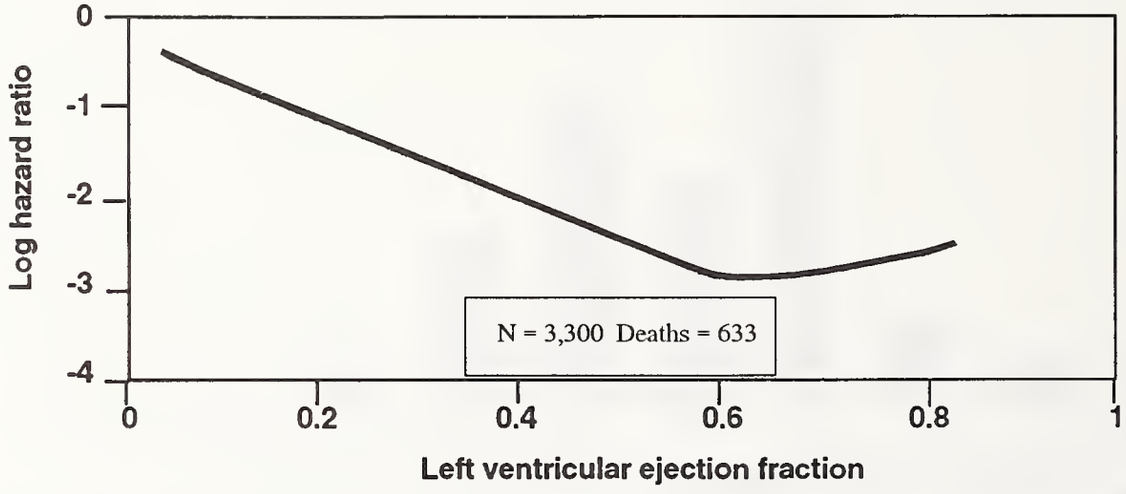
Functional form. The third problem in fitting a model is the functional form of the covariate. For example, is the incremental risk the same over the entire range of the covariate? Should x be used as measured or some function of x? Several investigators (Harrell and Lee, 1986; Harrell, Lee, and Pollock, 1988; Harrell, Pollock, and Lee, 1987; Stone and Koo, 1985; Durrelman and Simon, 1989; Sleeper and Harrington, 1990) described an approach in which splines are used to determine the shape of the function of the covariate. For instance, in the study by Califf, Harrell, Lee, and others (1989), the left ventricular ejection fraction (LVEF) for the population ranged from 6 to 91. A cubic spline fit to the data (Figure 2) indicated that an LVEF above 60 had no incremental benefit. By truncating the LVEF to 60, the model likelihood χ^2 was increased significantly. For more complex functional forms, each component of the spline fit is entered in the model. To test for an overall covariate effect, a χ^2 with k-1 degrees of freedom is computed where k is the number of knots (join points) required to model the effect. Linearity is tested with k-2 degrees of freedom.

Investigators (for example, Therneau, Grambasch, and Fleming, 1990; Schoenfeld, 1982; Pettitt and Daud, 1990) have proposed methodologies for examining the nature of the effects of covariables across time. The use of time-dependent covariates to model the covariables' effect over time is also possible with parametric survival models (Herndon and Harrell, 1989).

Number of covariates. The fourth problem is the number of covariates a given data set can support. It is well known—but seldom taken into account—that as the number of parameters estimated exceeds a given value, predictions on a new sample actually become worse. This problem is particularly acute with regard to estimation of absolute survival probabilities.

The most reliable predictor is one that always predicts the mean outcome. If the end users desire discrimination ability, some reliability has to be sacrificed. The question then becomes "How many parameters can one estimate and still have reasonable reliability?" In fact, if

Figure 2. Spline transformation for left ventricular ejection fraction



Source: Duke University Heart Center.

too many parameters are estimated, even the discrimination will worsen due to overfitting (Harrell, Lee, Califf, and others, 1984).

To demonstrate this effect, we did two simulation experiments involving fitting Cox models with p uniformly distributed (on [-0.5,0.5]) covariates, x, when p is varied from 1 to 29. In the first experiment, for each p, 10 samples of size n = 750 were drawn from an exponential distribution with true hazard $\lambda\exp(\underline{x\beta})$ where $\underline{x\beta}=x_1\beta_1+x_2\beta_2+\cdots+x_p\beta_p$, and $\beta_1=\beta_2=\cdots=\beta_p=0.25$. λ was chosen to yield an average 5-year survival of 0.75. The samples had an average of 185 deaths. For each sample, a Cox model was fitted to the p covariates, and the predicted 5-year survival was computed for all subjects. These estimates were compared with the true survival probabilities for the same subjects. Average absolute errors were averaged over the 750 subjects and over the 10 samples. A second experiment was done with n = 250, with an average of 64 deaths at 5 years.

In Figure 3, the solid lines represent results from the n = 750n experiment and the dashed lines represent results from the n = 250 experiment. Results for the "sick" (high-risk) patients are denoted with boxes (■). Lines without boxes denote results for the "average" patient. The curves for the average patients were constructed by averaging the error over the entire patient population. The curves for the sick patients were constructed by averaging over patients from the lower decile of predicted survival.

If the investigator wishes to limit the average error to some fixed value, say 0.05, there is an upper limit on the number of parameters to be estimated. For the average patient, in the n = 750 experiment, the expected survival error is monotonically increasing in p and exceeds 0.05 when p > 14. For the n = 250 experiment, the error reaches 0.05 when p > 5. Absolute survival error rises more steeply for the highest decile of predicted risk. Survival estimation is often crucial for these sickest patients. Here, the error exceeds 0.05 for p > 5 (n = 750) and for p > 3 (n = 250). Thus, when the number of parameters is large compared with the number of events in the sample, overfitting occurs. This overfitting has more effect for patients with very high or very low predicted survival, because overfitting causes predicted survival to shrink toward the mean survival. Harrell, Lee, Matchar, and Reichert (1985) proposed that there should be at least 10 events for every candidate covariate in order to guard against this overfitting. Other work has shown this applies almost equally in a stepwise variable selection where p is the number of candidate variables. We need to constantly remind ourselves of the penalty for creating very complex models.

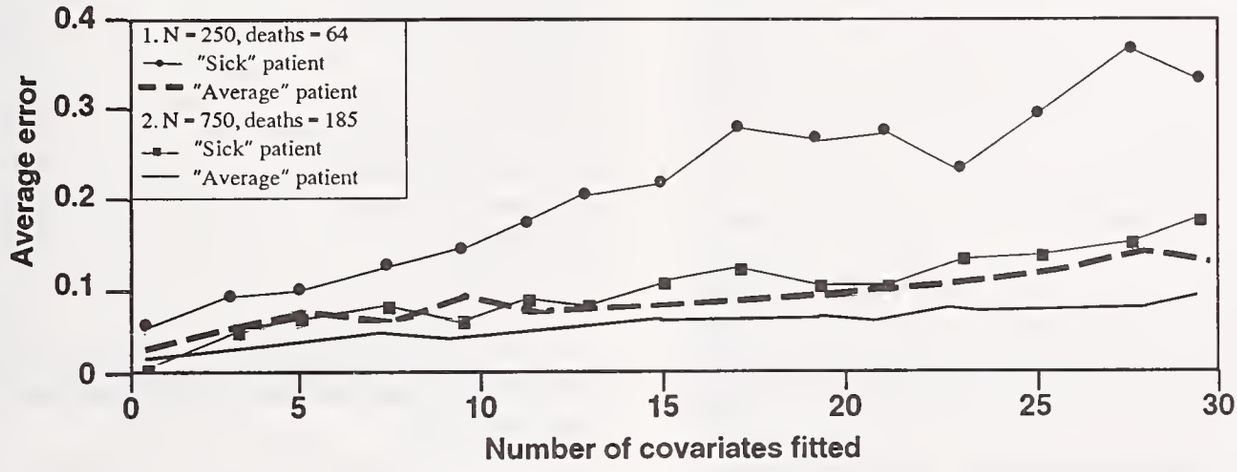
One final point is that covariate selection should be done with the physician and the statistician working as a team, each supplying their own particular expertise and knowledge to the enterprise.

Prediction

Broadly speaking, there are two functions of prediction—forecasting and hypothesis testing (see Bunge, 1979, for a general discussion). In survival analysis, forecasting is the deduction of future probabilities of survival for a given patient or population. An example of hypothesis testing is confirming or refuting the hypothesis that a specific covariate influences survival.

Naftel (in this volume) points out that survival models provide curves for each treatment to both the patient and the physician; these curves represent the most likely probability of survival at each time point. Survival prediction serves a number of worthwhile purposes.

Figure 3. Average error in Cox 5-year survival estimates



Source: Duke University Heart Center.

By comparing the curves over time, the patient and physician can make a more informed decision about treatment selection. If the patient's personal goal is to maximize the likelihood of short-term survival, a specific treatment may be chosen. If the patient's goal is to maximize long-term survival, an alternative treatment may be chosen.

On a larger scale, the prediction curves can provide information to public and private health care providers about the most likely survival curve for specific patient groups. Given the predicted survival curves, these agencies may recommend different treatments and reimbursement strategies for different patient groups.

Multiphase Model

The multiphase model is structurally more complex than other parametric models. The primary goal in its development was to model a wide variety of survival shapes and provide a means to determine how covariates influence various parts of the survival curve.

Strengths. The multiphase model can estimate the effects of covariates at different time segments along the survival curve. Table 2 lists covariates that influenced outcome in 2,967 coronary artery bypass graft patients (Califf, Harrell, Lee, and others, 1989). Smith, Harrell, Rankin, and others (1991) used the multiphase model to analyze these data. Table 2 lists these results along with those from an analysis using logistic regression for 30-day mortality and with results from the Cox model for long-term survival. A comparison of these three analyses shows a strength of the multiphase model.

The early phase mimics results of the logistic model. Both have the ability to estimate covariate effects when the covariate is influential for a short time after treatment is begun. It is possible that the initial hazard due to the treatment may last for shorter or longer periods of time depending on other patient characteristics. In such a case, the logistic model using fixed-length followup (in-hospital, 30-day, or 6-month mortality) may not be adequate to describe accurately the relevant covariates that affect the early treatment hazard. The multiphase model can estimate such effects more accurately, provided that patients are followed long enough to define the early phase accurately.

The Cox model identifies covariates that are found in at least one of the phases of the multiphase model. The Cox model "averages" over the entire hazard, whereas the multiphase model identifies covariates that dominate the hazard at particular times. Thus, the Cox model and the multiphase model tend to give consistent results overall, but the multiphase model yields additional information on when the covariates most influence survival. A modification of the Cox model can permit the partition of survival into separate phases but at the expense of considerable computational burden (Anderson and Senthilselvan, 1982).

Another strength of the model is that it is implemented as a procedure in SAS (Blackstone and Naftel, 1988), one of the most versatile, powerful, and widely used of the extant statistical analysis software packages. Thus, data input and results output are relatively straightforward.

Weaknesses. The learning curve for using this model is steep and daunting. In order to model some survival curves accurately, up to three different hazard functions must be specified with as many as nine shaping parameters (four in the early phase, one in the middle phase, four in the late phase) to be estimated. One successful strategy for estimating the hazard parameters is to include only one distribution and achieve the best fit, then fix the parameters and add a second phase; when the best fit is achieved, fix the parameters in the second phase and add the third phase. When parameters for this last phase have been estimated, constraints on all parameters are removed and the model is refitted to get a complete set of hazard parameter estimates. At each step, the significance of the phase can be assessed with the likelihood ratio test.

When all the phases have been identified, the hazard parameters are fixed, and the separate streams of covariates are added. When covariate selection is completed, all constraints on the hazard parameters are released and the model is run again to get final estimates for the hazard parameters, covariate effects, and the covariance matrix. This entire procedure can be long, onerous, and

Table 2. Comparison of covariates selected by three survival models

Covariate	Logistic 30 day death	Multiphase[1] early	Multiphase[1] late	Cox model
Earlier year of surgery	•	•	•	•
Lower ejection fraction	•	•	•	•
Greater extent of coronary artery disease	•	•	•	•
Older age	•	•	•	•
More conduction abnormality	•	•	•	•
History of hypertension	•	•	•	•
Greater anginal symptoms	•	•		•
Lower weight	•	•		•
Greater vascular disease			•	•
Presence of diabetes			•	•
Greater myocardial damage			•	•

[1]Multiphase analysis originally published by Smith, Harrell, Rankin, and others, 1991.

Note: A dot (•) indicates that covariate was significant ($p < .05$).

expensive. In addition, up to three streams of covariate effects may be modeled. Thus, the previously described problems of finding spuriously significant covariates are compounded greatly.

As currently implemented, there is no ability to test for equal covariate effects across phases during the stepwise covariate selection process. Only after covariate selection can a test be done on those covariates selected for two or more phases. If a covariate is selected for one phase but not for another, no test of equality of effect is possible unless the covariate is forced into the model. Furthermore, if a covariate has a similar effect over two or more phases, there is no facility for estimating a common effect. The next version of the model, however, will have this very useful facility (Naftel, personal communication, September 1991).

We recommend that the papers by Blackstone, Naftel, and Turner (1986), Bradley, Bradley, and Naftel (1984), Turner (1975), Turner, Hazelrig, and Blackstone (1982), and Turner and Pruitt (1978) be reviewed in order to understand the interrelationships among the distribution subfamilies.

Summary

Many models are available for use in assessing treatment effects, including the nonparametric models such as the Cox model, the parametric variations on the Cox model (Herndon and Harrell, 1989), and other parametric models (for instance, Bailey, 1988), as well as the multiphase model. These models may incorporate separate streams of covariates or use time-dependent covariates to assess treatment effect. Each of these models is under continual development and offers great promise for more accurate identification of important covariates.

No model or process completely solves all the problems outlined here, but the multiphase model is a new and powerful analytic technique for treatment comparison, covariate selection, and prediction and deserves further evaluation. It has been used successfully in a wide range of applications requiring many different hazard shapes.[2] The hallmark of any survival modeling strategy should be simplicity, and everyone engaged in this pursuit should carefully examine whether the problem and the quality and quantity of data at hand warrant the complexity of the model (or indeed, any model) being used. Also, the more complex the model, the more important validation becomes.

[2]The American College of Cardiology/American Heart Association Task Force Report (1991) has extensive examples of its use in studies of coronary artery disease, and an extensive bibliography.

[3]Pryor, D.B., Principal investigator, AHCPR grant no. HS06353-01, Outcome assessment program in ischemic heart disease. 1989.

Other issues that were not discussed by Naftel (in this volume) but that ultimately bear on the survival analysis problem include the use of treatment data not collected in a randomized clinical trial and the analysis of treatment data where treatment is confounded with data collection site. These issues have been addressed (Hlatky, Califf, Harrell, and others, 1988)[3], and continue to be studied.

The prediction process has some interesting ethical ramifications. The method permits an objective assessment of risk. The patient and the health care provider, however, looking at the same treatment survival curves may make two diametrically opposite choices. Conflicts may arise if the health care providers are unwilling to pay for the patient's choice. Thus, there is potential for conflict between two sets of ethics: those of the patient as an individual and those of society as a whole.

Survival analysis, with its ability to predict the benefit of one treatment over another, is only one part of the larger problem of how to allocate health care resources. The next step is to be able to predict the cost (either cost to the patient, cost to the health care provider, or in some sense, cost to society) with the same precision and accuracy that we can predict benefit. The modeling process could play a central role in the resolution of this important problem.

References

American College of Cardiology/American Heart Association Task Force Report. (1991). Guidelines and indications for coronary artery bypass graft surgery. *Journal of the American College of Cardiology*, 17(3), 543-589.

Anderson, J.A. and Senthilselvan, A. (1982). A two-step regression model for hazard functions. *Applied Statistics*, 31(1), 44-51.

Bailey, R.C. (1988). Some uses of a modified Makeham model to evaluate medical practice. *Journal of the Washington Academy of Sciences*, 78(4), 339-353.

Blackstone, E.H. and Naftel, D.C. (1988). *SAS technical report P-175, Changes and enhancements to the SAS system, Release 5.18, under OS and CMS*. Cary, NC: SAS Institute.

Blackstone, E.H., Naftel, D.C., and Turner, M.E., Jr. (1986). The decomposition of time-varying hazard into phases, each incorporating a separate stream of concomitant information. *Journal of the American Statistical Association*, 81(395), 615-624.

Bradley, D.H., Bradley, E.L., and Naftel, D.C. (1984). A generalized Gompertz-Rayleigh model as a survival distribution. *Mathematical Biosciences*, 70(2). 195-202.

Bunge, M. (1979). *Causality and modern science* (3rd rev.) (pp. 310). New York: Dover Publications.

Califf, R.M., Harrell, F.E., Jr., Lee, K.L., and others. (1989). The evolution of medical and surgical therapy for coronary artery disease: A 15 year perspective. *Journal of the American Medical Association*, 261, 2077-2086.

Califf, R.M., Mark, D.B., Harrell, F.E., Jr., and others. (1988). Importance of clinical measures of ischemia in the prognosis of patients with documented coronary artery disease. *Journal of the American College of Cardiology*, 11(1), 20-26.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, Ser B, 34(2), 187-220.

Durrelman, S. and Simon, R. (1989). Flexible regression models with cubic splines. *Statistics in Medicine*, 8, 551-561.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *CBMS-NSF regional conference series in applied mathematics,* Vol. 38. Philadelphia: Society for Industrial and Applied Mathematics.

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1), 36-48.

Feigl, P. and Zelen, M. (1965). Estimation of exponential survival probabilities with concomitant information. *Biometrics*, 21, 826-838.

Harrell, F.E., Jr. and Lee, K.L. (1986). Verifying assumptions of the Cox proportional hazards model. *Proceedings of the 11th Annual SAS User's Group International Conference* (pp. 823-828). Cary, NC: SAS Institute.

Harrell, F.E., Jr., Lee, K.L., Califf, R.M., and others. (1984). Regression modeling strategies for improved diagnostic prediction. *Statistics in Medicine*, 3, 143-152.

Harrell, F.E., Jr., Lee, K.L., Matchar, D.B., and Reichert, T.A. (1985). Regression models for prognostic prediction: Advantages, problems, and suggested solutions. *Cancer Treatment Reports*, 69(10), 1071-1077.

Harrell, F.E., Jr., Lee, K.L., and Pollock, B.G. (1988). Regression models in clinical studies: Determining relationships between predictors and response. *Journal of the National Cancer Institute*, 80(15), 10-13.

Harrell, F.E., Jr., Pollock, B.G., and Lee, K.L. (1987). Graphical methods in survival analysis. Proceedings of the 12th Annual *SAS User's Group International Conference* (pp. 1107-1115). Cary, NC: SAS Institute.

Herndon, J.E., II and Harrell, F.E., Jr. (1989). Examining the relationship between survival time and time dependent covariables using the PHSPLM procedure. *Proceedings of the 14th Annual SAS Users Group International Conference* (pp. 1318-1323). Cary, NC: SAS Institute.

Hlatky, M.A., Califf, R.M., Harrell, F.E., Jr., and others. (1988). Comparison of predictions based on observational data with the results of randomized controlled clinical trials of coronary artery bypass surgery. *Journal of the American College of Cardiology*, 11(2), 237-245.

Pettitt, A.N. and Daud, I.B. (1990). Investigating time dependence in Cox's proportional hazards model. *Applied Statistics*, 39(3), 313-329.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1), 239-241.

Sleeper, L.A. and Harrington, D.P. (1990). Regression splines in the Cox model with application to covariate effects in liver disease. *Journal of the American Statistical Association*, 85, 941-949.

Smith, L.R., Harrell, F.E., Jr., Rankin, J.S., and others. (1991). Determinants of early versus late cardiac death in patients undergoing coronary artery bypass graft surgery. *Circulation*, 84(Suppl. III), 245-250.

Stone, C.J. and Koo, C.Y. (1985). Additive splines in statistics. *Proceedings of the Statistical Computing Section* (pp. 45-48). Washington: American Statistical Association.

Thall, P.F. and Lachin, J.M. (1986). Assessment of stratum-covariate interactions in Cox's proportional hazards regression model. *Statistics in Medicine*, 5, 73-83.

Therneau, T.M., Grambasch, P.M., and Fleming, T.R. (1990). Martingale-based residuals for survival analysis. *Biometrika*, 77(1), 147-160.

Turner, M.E., Jr. (1975). Some classes of hit-theory models. *Mathematical Biosciences*, 23, 219-235.

Turner, M.E., Jr., Hazelrig, J.B., and Blackstone, E.H. (1982). Bounded survival. *Mathematical Biosciences*, 59, 33-46.

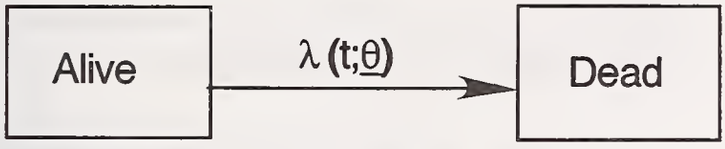
Turner, M.E., Jr. and Pruitt, K.M. (1978). A common basis for survival, growth, and autocatalysis. *Mathematical Biosciences*, 39, 113-123.

Appendix

Survival models. In general, survival models can be thought of as a special case of a two-state (alive, dead) Markov process (appendix figure). The transition rate from the alive state to the dead state is governed by the hazard function $\lambda(t;\underline{\Theta})$ where t is time and $\underline{\Theta}$ is a vector of parameters that govern the shape of the hazard function. The corresponding survival function is $S(t)= \exp\{-\Lambda(t,\underline{\Theta})\}$, where $\Lambda(t,\underline{\Theta})$ is the cumulative hazard up to time t. The central issue of survival analysis is to describe $\lambda(t;\underline{\Theta})$ and to determine what covariates modify $\lambda(t;\underline{\Theta})$. There are two general methods of incorporating covariates into the model. The first is to assume that one or more of the parameters, Θ, are linear combinations of the covariates, $\underline{x}$ (eg $\Theta_j=\underline{x}\underline{\beta}$). This

method was used early by Feigl and Zelen (1965) and more recently by Bailey (1988).

Appendix figure.



Another method is to multiplicatively scale $\lambda(t;\underline{\Theta})$ by $\exp(\underline{x}\underline{\beta})$ so that the adjusted hazard is $\lambda(t;\underline{x},\underline{\Theta}) = \lambda(t;\underline{\Theta}) \exp(\underline{x}\underline{\beta})$. This is the proportional hazards formulation. If a specific parametric form of the hazard function is assumed, the model is called parametric. If no specific function form is assumed, then the model is semi-parametric (Cox, 1972).

The multiphase model assumes the cumulative hazard to be the weighted sum of three hazards, where the weights are scaling functions of the covariates. One phase describes the early phase of survival. A second phase describes intermediate survival. The third phase describes late survival. Each of these three phases is present over all time but dominates survival during only one time interval. The total cumulative hazard function is the sum of the three phases; that is

$$\Lambda_T(t) = \mu_1(\underline{x}_1,\underline{\beta}_1) \cdot \Lambda_1(t,\underline{\Theta}_1) + \mu_2(\underline{x}_2,\underline{\beta}_2) \cdot \Lambda_2(t,\underline{\Theta}_2) + \mu_3(\underline{x}_3,\underline{\beta}_3) \cdot \Lambda_3(t,\underline{\Theta}_3), \Theta$$

where, for i=1(early), 2(middle), and 3(late), $\mu_i = \exp(\underline{x}_i\underline{\beta}_i)$ is the scaling parametric function of the covariates, $\underline{x}_i$=vector of covariates, $\underline{\beta}_i$=vector of regression coefficients, Λ_i=cumulative hazard function, $\underline{\Theta}_i$=vector of parameters governing the shape of the cumulative hazard, and t=time to failure. The specific functional forms of Λ_i are given by Blackstone, Naftel, and Turner (1986). The survival function is given by $S(t)=\exp\{-\Lambda_T(t)\}$. The model permits the description of a survival function using any one, two, or all three of these phases.

The form of the scaling parametric functions, μ_i, is equivalent to a proportional hazards assumption for each phase of the model. Because the model can describe each phase of the survival curve by a different set of covariates, the separate phases of the survival curve can be adjusted independently.

Episode of Care Framework: Utility for Medical Effectiveness Research

Donald M. Steinwachs, Ph.D.

Introduction

The effectiveness of medical care has been a focal issue for the field of health services research since its inception. Unlike efficacy research that evaluates specific treatments under ideal conditions, effectiveness research addresses the impact of health service interventions under conditions of routine practice in the community. Since clinical practice is continually changing in response to new knowledge and changing needs for health services, the challenge for effectiveness research is to provide timely information on who benefits from which diagnostic and treatment services and how services can be provided to assure maximum health benefits—that is, the best possible patient outcomes.

Over the past three decades, we have learned that effectiveness can be compromised by many factors: delay in timely diagnosis and treatment of a condition, inappropriate choice of treatment, failures in compliance with treatment, and failures in adequate followup care. To understand the sources of problems that reduce the effectiveness of care, we have examined in depth the characteristics of the health care system, including financing, organization, manpower, and technology and their relationship to the characteristics of populations served by these systems.

These efforts have utilized many methodologies and units of observation, including the community (Kessner, Kalk, and Singer, 1973), the patient, the provider, and the specific service or treatment received. A unit of observation that has received relatively little attention is the "episode of care," yet the episode of care provides the conceptual framework for much of the current research on effectiveness.

Dr. Steinwachs is Director, Health Services Research and Development Center, and Professor, Department of Health Policy and Management, Johns Hopkins School of Hygiene and Public Health.

What is an episode of care? In one of the first studies to use an episode of care, the episode was defined as "one or more medical services received by an individual during a period of relatively continuous contact with one or more providers of service in relation to a particular medical problem or situation" (Solon, Feeney, Jones, and others, 1967).

It is important to keep in mind that conceptually the episode of care has a relationship to an episode of illness, but there may be little information available by which to define this relationship precisely. Indeed, it is critical to keep in mind that medical care utilization occurs in the absence of an episode of illness, and more importantly, many episodes of illness do not lead to any utilization of medical care. Thus, we should be cautious whenever we use the episode of care as a proxy for the episode of illness.

Why is it relevant for medical effectiveness research? The relevance of the episode of care framework for medical effectiveness research depends, in part, on the research strategy underlying the development of effectiveness information. If the strategy involves moving from studies that attempt to identify the sources and magnitude of problems in effectiveness to studies that investigate the structure and processes of care that contribute to variations in effectiveness, then the episode of care is central. It is the episode of care that precisely defines the process of care. Where this process occurs identifies the system(s) of care whose characteristics may affect outcomes. Thus, the episode framework provides the key independent variables related to the structure and process of care that are needed to examine variations in patient outcomes and effectiveness.

The potential utility of the episode of care for describing and analyzing the processes of diagnosis, treatment, and outcome assessment has been repeatedly advanced; yet the episode of care framework has had relatively limited application. One explanation may be that it is al-

ways assumed to be the starting point for medical effectiveness research and does not need to be formally acknowledged. In concept this may be true, but in actual practice many key aspects of the episode of care appear to be ignored.

For example, research on the effectiveness of specific surgical procedures—such as prostatectomy or cataract removal—frequently starts with the surgery and links information on postsurgical outcomes. It is not always clear that the research examines comprehensively the presurgical care processes that lead to the diagnosis and treatment. It is in the presurgical phase that the diagnosis is reached and consideration may be given to nonsurgical treatments. As we know, only a subset of all presenting patients ultimately experience surgery, occasionally due to diagnostic errors, but usually due to the choice of no treatment or alternative nonsurgical treatment. This is a potential source of bias. Focusing on the outcomes of only one treatment approach (such as surgery) ignores the outcomes from other treatments or no active treatment.

One reason for giving priority to the use of an episode of care framework is to quantify the magnitude of the "treated group" bias. Also, it can be used to identify comparison groups for contrasting the outcomes of alternative treatments or of no active treatment. Essentially, the episode framework shifts the starting point of the analysis from the selected treatment of interest to the presentation of a patient with a specific problem. Starting with the presentation of the patient is critical if we are to begin to learn how to influence the treatment process in the direction of increasing the overall effectiveness of medical care.

Why episode of care analyses are not done more frequently? The centrality of the episode of care to medical effectiveness research is reasonably clear; yet the development of information on the characteristics of the episode of care is frequently inconsistent or absent. Several barriers to the routine use of the episode of care as a unit of analysis or as a defining framework for the process of care include:

1. Lack of available criteria for defining episodes of care that have been tested for validity and reliability;

2. Limitations of existing routine data systems, including claims data, regarding information needed to construct episodes of care;

3. Uncertainty regarding the application of an episode of care methodology to chronic, long-term conditions; and

4. Lack of standard summary measures to characterize an episode—for example, duration, intensity, utilization, cost, and outcomes.

For a comprehensive discussion of episode of care definitions, issues, and applications through the 1970s and early 1980s, the review by Hornbrook, Hurtado, and Johnson (1985) should be carefully read. Progress is being made in addressing these potential barriers; and with continuing investments in methodological research, it should be possible to effectively deal with all of them.

Applications of the Episode of Care

The episode of care has been used to examine issues of provider productivity, cost of care, patterns of utilization, and quality of care. Although the range of applications has been broad, the numbers of applications have been relatively few.

Utilization studies. One of the compelling reasons for developing episode of care information is to identify the extent and type of utilization associated with treating a specific episode of illness. Measures of utilization may include numbers of visits, hospital admissions, prescribed medications, laboratory tests, home health services, and other types of health care. One application of episode-based utilization data was in the Graduate Medical Education National Advisory Committee (GMENAC) report (1980) on physician requirements. The Committee used a modified epidemiologic model of incidence and prevalence of disease, percent needing care, and for those needing care, numbers of visits per episode (or per year for chronic conditions). In comparing expert opinion with actual episode utilization experience in health maintenance organizations (HMOs), Weiner, Steinwachs, Shapiro, and others (1987) found the GMENAC estimates of utilization per episode to be somewhat higher than found in actual practice, contributing to higher estimates of physician requirements than current HMO practice patterns would suggest are needed (Steinwachs, Weiner, Shapiro, and others, 1986).

An episode of care framework was used to examine the impact of cost sharing in the RAND Health Insurance Experiment (Keeler, Buchanan, Rolph, and others, 1988; Keeler and Rolph, 1982). These analyses provided important insights into the overall observations that as cost sharing increased, utilization declined. Keeler and colleagues (1988) found that the decline in utilization was associated with fewer treated episodes of care and not with fewer services being provided during the episode of treatment. Cost sharing through deductibles and copayments discouraged individuals from seeking care for specific problems but apparently did not discourage them from receiving the usual array of services once care was sought.

Other applications of episode utilization models are being developed in models of managed ambulatory care. Comparisons are made between actual use and criteria specifying upper bounds on use per episode. Donald Harrington has developed a proprietary system used by some of the major insurers that had its early beginnings in the development of a utilization review system for the San Joaquin Medical Care Foundation in California in the 1970s (Shapiro, Steinwachs, Skinner, and Mushlin, 1976). This type of utilization review—and in some cases precertification process—is expected to become more common as managed care is extended into the ambulatory care setting (Institute of Medicine, 1989).

Other studies that have examined utilization patterns using episodes include Fink (1973); German, Skinner, Shapiro, and Salkever (1976); Keeler, Wells, Manning, and others (1986); Keeler and Rolph (1988); Kessler, Steinwachs, and Hankin (1980); Moscovice (1977); Salkever (1976); Solon, Rigg, Jones, and others (1969); and Young and Fisher (1980).

Cost studies. A relatively unusual perspective on trends in health care costs was developed using episodes of care (Scitovsky and McCall, 1976). They showed how the trends in episode utilization have driven costs as well as raised prices for specific services. These trend analyses helped sharpen issues regarding changes in medical practice and what is gained by more expensive treatment of many common conditions.

Two other studies used an episode of care centered around a hospitalization to examine the impact of hospital payment policies on reducing costs. Rupp, Steinwachs, and Salkever (1984) assessed the impact of per case hospital payment in Maryland on psychiatric episodes of care for Medicare beneficiaries as compared with more traditional per service payment. More recently, Zawistowich (1989) examined the impact on psychiatric episode costs of day limits imposed by Maryland Medicaid. She compared episodes centered on a hospitalization over three time periods to determine if a reduction in the payment limit from 18 days to 16 days for psychoses saved Medicaid money. These studies illustrate the utility of episode analyses for assessing the impact of hospital payment policies on total costs, including inpatient hospital and physician costs, preadmission and postdischarge ambulatory care, and any hospital readmissions within the episode of care time period. In both cases, the investigators used episode analyses to determine if changes in hospital payment resulted in a substitution of ambulatory care for inpatient care or in more frequent hospitalizations.

Effectiveness studies. Relatively few effectiveness studies have used formal episode of care analyses, although as discussed above, the episode of care is the conceptual underpinning to many effectiveness studies. One example of the use of the episode model in ambulatory care was the Experimental Medical Care Review Organization's (EMCRO's) demonstration program at the Columbia Medical Plan in which a quality assurance model was tested with visit data linked to patient-reported outcomes (Mushlin and Appel, 1980). From this study came additional studies examining the effectiveness of nurse practitioners as compared with physicians and their comparative cost (Salkever, Skinner, Steinwachs, and Katz, 1982). Other investigations include those of Lohr and Brook (1980), who studied the quality of care in episodes of respiratory illness, and Lasdon and Sigmann (1977), who examined the cost and effectiveness of hypertensive care using episodes. This is a natural framework for the analysis of variations in effectiveness.

Summary. Overall, the literature explicitly concerning use of episodes of care to examine utilization, cost, and effectiveness of health care is relatively limited. This appears to reflect the complexity in developing episode of care information and probably uncertainties regarding how to summarize utilization and cost for episodes of care. Another contributing factor may be the limited information available in most routine data sources (for example, claims data) that can be used to link individual services into episodes of care and to assess outcomes of care. Even though only a small number of studies have used the episode of care, it is apparent that it can provide unique insights into factors affecting care-seeking, patterns of practice, outcomes of care, and costs.

Methodological Lessons from Previous Research Using Episodes

Several important lessons have been learned from previous efforts. These relate to the development of episode of care data and to the measurement of episode characteristics. As has been demonstrated, explicit criteria can be used to define episodes; and in a few cases, these have been validated and tested for robustness. The episode has been shown to provide important and unique insights into the pattern of care, the roles of different types of health professionals and technologies, and the cost of care. Few applications have been made to the care of chronic conditions, but these are promising. The current challenges are to develop improved methods for addressing issues of severity and outcomes of care for episodes.

Construction of episode of care data. Solon (1967) and later Scitovsky and McCall (1976) constructed episodes of care through the review of medical records and

billing data in order to describe utilization during episodes of care and, in the Scitovsky and McCall study, to track total costs (charges) for the care provided. With the work of Moscovice (1977) and Steinwachs and Yaffe (1974), efforts were made to develop explicit criteria for linking services into episodes of care using ambulatory care information systems. The value of explicit criteria related both to reproducibility and to the opportunity to use computers with large data sets to construct episodes of care.

It was not until the work of Kessler, Steinwachs, and Hankin (1980) that an independent validation of the application of explicit criteria for episode development occurred. This study examined episodes of care for the treatment of mental illnesses, conditions that have both acute and frequently chronic or recurrent aspects. A sample of episodes constructed by clinicians from medical records was compared with those generated by the computer using claims (encounter) data and explicit criteria. Kessler and coworkers found that an 8-week criterion for the maximum interval between visits in the same episode had a specificity of 85 percent but a sensitivity of only 41 percent. The lower-than-expected sensitivity reflected breaks in the regular pattern of visits of longer than 8 weeks which was considered by the provider to represent continuing treatment and not a new episode.

Related work done by Roos, Cageorge, Austen, and Lohr (1985) used claims data to identify postsurgical complications. The criteria that linked postdischarge care with the index surgery represent the partial construction of the episode of care with emphasis on indicators of outcome. The investigators validated the criteria using an independent review of medical records by clinicians and found high levels of agreement.

Overall, this work demonstrates that explicit criteria can be used and, equally importantly, can be validated by clinical assessment. An interesting question is whether having patients, instead of providers, define episodes of care would lead to the same results. It is almost certain that episodes defined by patients will differ but possibly not very substantially. Patient validation should be tested.

Robustness of criteria for constructing episodes. One approach used in several studies relying on episode of care data is to derive criteria that have face validity and are robust (Keeler and Rolph, 1982; Keeler, Wells, Manning, and others, 1986; Weiner, Steinwachs, Shapiro, and others, 1987). Through the use of this strategy, criteria for linking services into episodes are defined in terms of ranges—from restrictive to liberal. Data are analyzed to determine if the construction of episodes is sensitive to where within the range the criterion is chosen. In the best of circumstances, little or no sensitivity is found. Alternatively, search procedures may be used to find the least sensitive criterion value. The caution, however, is that robust criteria need not necessarily be valid.

Even so, it is important to examine the robustness of any criterion used, including those independently validated. Valid criteria that are not robust may contribute to findings of "no significant difference." In general, one should be able to develop explicit criteria that meet both tests; that is, they should be independently validated and should be demonstrated to be robust in their application.

Classification issues and potential biases. Several common classification issues are related to the construction of episodes of care. These will be briefly reviewed and their implications discussed.

When does the episode begin? Does it begin when the patient presents with a complaint that ultimately resolves into the diagnosis of interest, or does it begin with the first visit or hospitalization with the specific diagnosis? Most applications of episode of care methods have started with the diagnosis; however, some investigators have attempted to include recent prior visits with related but ill-defined symptoms (Steinwachs and Yaffe, 1974). Conceptually, one would like to begin with a presenting complaint, but data systems do not generally capture this information. The one exception is the National Ambulatory Medical Care Survey of the National Center for Health Statistics, but this data source describes individual visits and cannot be used to construct episodes.

The bias introduced by starting with the diagnosis is of two types. First, the beginning of the episode may be inaccurately identified. Second and more importantly is the possibility that individuals will have presented for care who were not diagnosed, or were inaccurately diagnosed, and will be totally missed by a failure to examine the presenting problem in defining the episode. The assumption is that this is an issue in many of the medical effectiveness research efforts. For example, in examining effectiveness of treatment for acute myocardial infarction, does one start with all those presenting with a symptom (for instance, chest pain) or a diagnosis? In general, we do not know how important the diagnostic bias is, but in some areas, it is easier to assess (there is considerable literature on outcomes of cases presenting with chest pain).

Another type of diagnostic bias is the presumption that the diagnosis recorded is meant to be definitive. For patients who were hospitalized, it is generally assumed that the listed discharge diagnoses are intended to be the provider's definitive statement. In ambulatory care, however, diagnoses are frequently not definitive—they may be statements of conditions to be "ruled out" or un-

certain but educated guesses (Steinwachs and Yaffe, 1978). Assuming these diagnoses are valid statements of what the provider was treating may lead to the over-identification of episodes of care; all "rule-out urinary tract infection" diagnoses may be counted as episodes of "urinary tract infection" when some proportion was not. Concern with this type of bias has led to several types of efforts to develop confirmatory information. For urinary tract infection, is there a prescription filled for an appropriate antibiotic, or for chronic conditions such as hypertension, is there more than one visit with a diagnosis (Steinwachs and Yaffe, 1978)? We have found that false positives are common for chronic conditions, but the occurrence of at least two visits with the diagnosis eliminates almost all false positives. Unfortunately, this strategy is not directly applicable to acute conditions where there may be no followup visit if the treatment is successful.

When does the episode end? The definition—the end occurs when all care related to the illness episode is complete—should include complications but may or may not include recurrences. If the recurrence is simply failure to treat the condition successfully, then it should be included. If the recurrence is another separate episode of illness, then it should be separate. Unfortunately, there is generally no way of being certain. Usually, time-interval criteria that have face validity to providers are used. The interval is chosen such that visits beyond this interval are thought to represent a "new" illness and not persistence of the previous illness. It is important to test the robustness of the criteria—that is, whether the identification of recurrent episodes sensitive to small variations in the time interval is used.

Can episodes be applied to chronic conditions? There is no consensus on this issue. It has been proposed that the long-term care of chronic conditions be divided into discrete episodes of treatment, including the initial episode when diagnosed, periodic management episodes, and any acute episodes associated with the chronic condition (Lasdon and Sigmann, 1977). Alternatively, some have taken chronic conditions and identified the initial diagnostic phase, ongoing management, and acute problems as parts of a single, possibly lifetime, episode. The approach used by Keeler and Rolph (1982) linked all routine chronic management services into one long episode; but if an acute problem occurred associated with the chronic condition, this was put into a separate episode of care. A criterion for choosing a preferred method for constructing episodes for chronic conditions is to begin with the measures of outcomes to be applied and select a method for describing prior utilization that describes relevant prior processes of care.

In general, the characterization of the process of care needs to be relevant to the objectives of the inquiry—in this case, effectiveness assessment. Effectiveness of care for chronic conditions may be measured by intermediate outcomes (such as blood pressure control for hypertension), long-term outcomes (such as reduction in stroke mortality in hypertensives), and avoidance of complications (such as potassium depletion from diuretic medications or hypotension in elderly patients). Each of these measures of outcome might benefit from a somewhat different aggregation of prior utilization into either episodes or time periods reflecting an understanding of the natural history of the disease and the effects of treatment. This area needs further methodological work to develop and test guidelines for the analysis of patterns of care and outcomes for chronic conditions.

How do we account for comorbidities in the episode of care? The problem of comorbidities being present is common to all analyses of medical effectiveness and is no different when episodes of care are used (Hornbrook, Hurtado, and Johnson, 1985). One must identify and statistically account for comorbidities in drawing conclusions regarding treatment and outcomes. An issue is whether none, some, or all of the utilization for significant comorbidity problems should be included in the episode of care. Little work has been done to address this question empirically. Reasonable options include using sensitivity analysis to examine the consequences of alternative inclusion criteria and clinical reasonableness to decide if the treatment of comorbidities is so related to the condition of interest that they are practically inseparable. These issues will assume even greater importance when the costs of providing care are examined in conjunction with the outcomes of care.

How do we measure the outcomes of episodes of care? This is certainly one of the central issues for medical effectiveness research. Are utilization-based measures of outcome valid and useful? Do we need routinely to capture other measures of outcome, for example, patient self-report? How do we link the episode of care and outcome information to provide insights into effectiveness?

There is now extensive experience with utilization-based measures of outcome but relatively little information on their validity or appropriate interpretation. Hospital readmissions that are judged to be potentially preventable have been used to evaluate the effects of hospital rate setting (Coelen and Sullivan, 1981). Another quality indicator is readmissions for complications related to surgery (Roos, Cageorge, Austen, and Lohr, 1985). It is important to recognize, however, that readmissions are part of the care pattern for many condi-

tions, for example, cancer therapy. In regard to those readmissions thought to be potentially preventable, there is little rigorous evidence regarding the extent of preventability under optimal care. Furthermore, readmission patterns are changing as we expand the range of ambulatory care services. Even though selected categories of hospital readmissions may be an important indicator of potentially poor quality care, the interpretation of changing readmission rates is frequently uncertain.

Another concern that arises when utilization-based indicators of outcome are relied on is the danger of interpreting the absence of readmission as a good outcome. This is similar to the issue of interpreting the absence of a malpractice claim as a measure of good care. In a recent study, roughly one in eight cases of medical negligence was estimated to result in a malpractice claim, and many malpractice claims were shown to have no merit (Localio, Lawthers, Brennan, and others, 1991). Thus, the absence of readmission is a relatively poor discriminator of outcomes.

For this reason, among others, Ellwood (1988) suggested that we should implement a system of outcome management. This system would involve the routine collection of information on patient outcomes of care. This concept is being tested in a variety of organized provider and insurer settings. Central to outcomes management is the linkage of patient outcomes to the process of care. The conceptual underpinnings come back to linking episodes of care and their outcomes through patient surveys and other methods that go beyond utilization-based measures.

When should outcomes of an episode of care be assessed? A fundamental issue concerns the timing of outcome measurements. To my knowledge, little attention has been paid to the question of when to measure patient outcomes. Hospital mortality is frequently calculated using death within 30 days of discharge, and readmissions are examined at intervals up to 1 year. The sensitivity of these measures to the use of different time intervals can be easily tested, since they are based on routinely available records. Patient-reported outcomes (Stewart, Greenfield, Hays, and others, 1989; Tarlov, Ware, Greenfield, and others, 1989) require investments in new data collection, and there has been less opportunity to compare different time intervals for different conditions.

Medical care is expected to have multiple benefits for the patient and to reduce risk of complications, anxiety, pain/discomfort, activity limitation, duration of symptoms, and risk of long-term disability. These potential benefits may be observable at different points in the medical care process, and thus, the timing of the "outcome" assessment can be expected to affect their detection. If outcomes are assessed early in the episode of care, we should have an opportunity to measure variations in duration and intensity of pain, anxiety, and activity limitations. If outcomes are assessed substantially after the episode of care, we will be able to measure long-term disability and late complications. Indeed, we are probably interested in both. The timing of single or multiple assessments is a critical methodological issue in identifying and assessing benefits/risks of medical care.

A related technical issue concerns quantification of the time interval between onset and outcome assessment. In concept, the time interval should be measured from when the episode of illness began and not when care was first sought and received, but generally, we will not know when this occurs. An alternative is to standardize the interval for measuring outcomes based on when the first service occurred, adjusted for the severity of the condition. This assumes that severity is a good proxy for delay in care seeking, which is probably only partially true. Yet this may be the best current method for standardizing outcome measurements obtained at different intervals from the onset of the episode of illness.

Summary. A variety of methodological problems must be faced in the construction and analysis of episodes of care. Essentially, these problems are inherent in most studies of medical effectiveness. Reasonable solutions are available to address these problems, but these solutions have not been the subject of rigorous methodological testing. As a result, there may be unrecognized biases, and certainly, the best solutions for these problems have not been clearly demonstrated. An issue that merits the highest priority for methodological research is to determine the appropriate timing of outcome measurements for different categories of conditions and types of outcomes. To do this will require the testing of measures of delay in seeking care (for example, provider or patient report, disease staging, severity scoring). With better estimates of the interval between onset of illness and outcome measurement, we should be able to determine more accurately the benefits and risks of alternative patterns of care.

Enhancing the Utility of Episode Data

The utility of the episode of care framework for effectiveness studies depends on its ease of application and its ease of interpretation. We have discussed many of these issues already and could conclude that episodes of care methods are not yet easy to apply or to interpret. Several steps can be taken to overcome these barriers. These include:

1. Developing standard terminology and definitions for episodes of care measures;

2. Refining methods for assessing the timeliness of the initiation of an episode of care;

3. Refining methods for assessing the severity and case-mix characteristics of episodes; and

4. Testing methods for routinely linking episodes with patient reported outcomes of care.

Characteristics of episodes. There is reasonable agreement on key characteristics of hospital discharges and ambulatory visits but not on episodes of care. The National Committee on Vital and Health Statistics took the initiative many years ago to define minimal data sets and establish definitions for basic categories of medical utilization (Department of Health and Human Services, 1981). This work needs to be extended to episodes of care.

Characteristics of episodes that should be defined and standardized include duration, cost, providers (for example, initial, principal), type of episode (ambulatory only, hospitalization, long-term care), type of condition (acute episodic, chronic, preventive), and type of criterion used in creating the episode (time interval, diagnosis, presenting complaint, test results, or other findings). By developing definitions and minimal sets of descriptors, we could significantly advance the application of episode methods (Sytema, Giel, and ten Horn, 1989).

Timeliness of care. The relationship of the episode of care to the underlying illness or other reason for seeking care is important. One way to measure this relationship is to understand the provider's assessment of the timeliness with which care is sought for the presenting problem. One study examining this issue found that in adults presenting to internal medicine, 10 percent of initial visits were rated as not needing care, 69 percent were rated timely, and 21 percent were rated as delayed (Steinwachs and Mushlin, 1978). In general, those falling into these groups were found to differ in terms of presenting problem status and outcomes measured at 3 weeks. Thus, it appeared that variations in provider-reported timeliness were related to variations in outcomes and effectiveness.

Other approaches to the measurement of timeliness include the development of disease-staging methods for specific conditions—for example, the staging of malignant tumors. More general disease-staging methods have been developed for the classification of hospitalized cases (Gonnella, Hornbrook, and Louis, 1984) but not for the broad spectrum of ambulatory care. This may represent a modest extension of previous work. The principal limitation is probably the lack of available routine data sources describing the full range of conditions and symptoms presented by the patient in the ambulatory care setting.

Severity and case mix. Substantial advances have been made in the measurement of case mix and severity for inpatient care (Fetter, Youngoo, Freeman, and others, 1980; Fries and Cooney, 1985; Horn and Horn, 1986; Iezzoni and Moskowitz, 1988; Knaus, Draper, Wagner, and Zimmerman, 1985; Young, 1984) and more recently for ambulatory care (Fetter, Averill, Liechtenstein, and others, 1984; Starfield, Weiner, Mumford, and Steinwachs, 1991; Weiner, Starfield, Steinwachs, and Mumford, 1991). None of this work has been explicitly applied to episodes of care. Specifically, one would like to measure the severity of the presenting condition at the beginning of the episode, not just when hospitalization occurs. Severity may be related to timeliness and can be expected to influence the duration, costs, and outcomes of the episode.

Also, there is a need to be able to adjust for case mix in the episode, that is, the mix of comorbidities present. It may be possible to use ambulatory care groups (Starfield, Weiner, Mumford, and Steinwachs, 1991) to do this, but this application has not been tested. Ambulatory care groups provide a classification predictive of overall ambulatory utilization based on the person's total mix of morbidities and demographic characteristics. This measure is usually applied to a 1-year period but might be applied to episodes of care by adjusting the expected utilization for each diagnostic cluster.

Linkage to outcome information. We have substantial experience with the linkage of utilization-based indicators of outcome (for example, complications leading to utilization), as well as mortality. They have many limitations, the principal one being that they measure some but not all of the negative outcomes and fail to measure any of the positive outcomes of care. If we apply an episode of care methodology, we may be able to add new utilization-based measures of outcome, including the duration of the episode of care and possibly the intensity of intervention required. Even so, this will not address the fundamental limitation of the absence of information on the full spectrum of health outcomes (Tarlov, Ware, Greenfield, and others, 1989).

The New England Medical Center, InterStudy, and others have recently begun to test the routine collection of information on patient-reported outcomes; these activities can be expected to open a new era in outcome assessment and the opportunity for linkage to episodes of care data. Currently, the emphasis is on identifying individuals at an early point in the episode of care and assessing presenting status and linking this to followup assessments to measure outcome status. This information is usually augmented with provider or record-based

information on characteristics of the condition, prior history, and current treatment. In organized care settings—for instance, hospitals and group practice HMOs—this methodology appears feasible. Implementation across a range of different types of settings with community providers is only beginning to be tested. The ultimate success will be the capacity to integrate outcomes assessment with existing management (claims) information systems to provide ongoing information regarding the relationship of diagnosis, treatment, and outcomes of care.

Application Toward Resolving Issues of Effectiveness

The principal rationale for making substantial investments to refine and improve episode of care methods is to provide better information regarding factors contributing to variations in effectiveness. Two areas in which episodes of care should provide unique insights are in identifying process-of-care correlates for outcomes and identifying where in the process of care to intervene to improve effectiveness.

Identifying correlates of outcome. In an epidemiologic approach, one thinks of risk factors for poor outcomes that can include personal and environmental characteristics, plus the characteristics of the medical care provided. Episode data can assist in probing into several of these avenues.

Incidence of treated episodes. With the use of episodes, the treated incidence of a condition can be estimated and compared with epidemiologic data on disease incidence to estimate extent of undertreatment (Steinwachs, Weiner, and Shapiro, 1989). One source of small area variations appears to be the lack of any treatment for some conditions in contrast to receiving some treatment that may involve surgical or medical care. We need to better understand the extent of undertreatment and its implications for outcomes if we are to improve the effectiveness of the health care system.

Access to care. Episodes—in particular self-reported episodes of illness and information pertaining to whether care was sought—have been used to examine access (Aday, Andersen, and Fleming, 1980; German, Skinner, Shapiro, and Salkever, 1976). If information systems based on encounter or claims data were used instead, problems in access could be identified by lower than expected incidence of treated episodes, initial visits for episodes to locations indicative of access problems (for example, emergency room use for routine problems), and measures of timeliness or severity that indicate late care seeking. We may find that many of the problems leading to poor outcomes have as their underlying cause delays in receiving care due to access problems and inappropriate care seeking.

Patterns of practice. The ability to characterize and classify patterns of practice is still in its early stage of development. There is a growing literature on some characteristics, including continuity of care, but there are few applications in which continuity has been measured and its benefits evaluated (Black, 1990; Starfield, Simborg, Horn, and Yourtee, 1976). Work on appropriate measures to describe patterns of practice—including continuity, the use of referrals and specialists, the use of diagnostic and monitoring tests, prescribing practices, the intensity and duration of treatment, the use of surgery versus other treatments, and the costliness of the practice pattern—is urgently needed. These key independent variables are expected to be related to outcomes and to assist in identifying inefficiencies.

Summary measures as suggested above will not be sufficient alone. They will need to be augmented by information on specific practices and treatments as is currently being done in medical effectiveness studies. Without the more general measures of patterns of practice, however, we will not be able to answer the critical question as to whether providers practice differently for different conditions or have a relatively consistent pattern of practice across all conditions. The nature of interventions to modify patterns of practice will be largely shaped by the answer to this question.

Compliance with provider recommendations. According to providers, a common cause of poor outcomes is poor patient compliance with medications and followup care. The episode provides a framework for examining provider followup, recommendations, and patient compliance. The specificity of this inquiry will depend on available information (for example, disposition at the time of a visit, prescriptions filled). In previous research, patients demonstrated relatively poor compliance; but compliance improves as patients get older, and females may comply better than males (Steinwachs, 1979). Some noncompliance may simply reflect an improving condition and be associated with good outcomes, whereas in other cases, it may contribute to delays in recovery and complications. Sorting out the importance of noncompliance for patient outcomes should be a high priority for effectiveness research and one that can be facilitated using episodes of care.

Implications for practice guidelines. Under the auspices of the Agency for Health Care Policy and Research, guidelines are currently being developed, and the dissemination process is underway. The development of guidelines involves integrating findings in the published literature and newly emerging evidence regarding effectiveness, through a process of expert con-

sensus, to define guidelines for the best practice of medicine. The question here is, "What can be added to this process by having information on episodes of care?"

Identifying practices associated with better outcomes. As an integrating framework, the episode of care can be used to identify general characteristics of provider practice patterns associated with better outcomes. Episode information may more clearly demonstrate whether provider continuity, specialty consultation, intensive workups, and patient compliance are associated with better or worse outcomes. This information can help shape what providers, patients, insurers, regulators, and employers are told each can do to improve the effectiveness of medical care.

Comparing cost and effectiveness of alternative approaches to diagnosis and treatment. As discussed previously, one inherent limitation in any study of effectiveness that begins with a specific treatment and examines previous and subsequent care is bias due to selectivity. The episode of care framework starts with the initiation of care for a problem and follows all identified individuals independent of the specific treatment. This makes it possible to identify the extent of bias in any given treatment subgroup and provides an opportunity to compare alternative treatments, adjusting for severity and other relevant prognostic factors. When this comparison is made, the episode provides information on total costs and duration that can be linked to outcomes. Thus, it provides the essential framework for examining cost and effectiveness that is central to making guideline recommendations.

Summary. Large data bases that include detailed information on the provision of services to individuals over time provide an opportunity to examine more comprehensively and rigorously the effectiveness, productivity, and efficiency of our health care system. Linking this information with patient-reported outcomes, even on a sample basis, will greatly enhance what can be learned regarding the effectiveness and costs of care provided. The episode of care provides one natural framework for this effort, and one that should substantially extend our understanding of the complex interactions between the patient and the health care system and how these influence patient outcomes. In promoting its application, investments will need to be made to:

1. Develop standardized criteria for aggregating services into episodes of care;
2. Define common terminology and summary measures of episodes;
3. Promote the adoption of critical data items in routine claims processing and managed care data systems; and
4. Provide tested guidelines for linking patient-reported outcomes information to the episode of care.

Making these investments will not only serve the field of medical effectiveness but also will provide enhanced capabilities for research into productivity and approaches for controlling the costs of care.

References

Aday, L.A., Andersen, R., and Fleming, G.V. (1980). *Health in the US: Equitable for whom?* Beverly Hills, CA: Sage Publications.

Black, C. (1990). *Population-based measurement of primary care to study variations in care received by the elderly.* Unpublished dissertation, The Johns Hopkins University.

Coelen, C. and Sullivan, D. (1981). An analysis of the effects of prospective reimbursement programs on hospital expenditures. *Health Care Financing Review*, 2(3), 1-40.

Department of Health and Human Services, National Committee on Vital and Health Statistics. (1981). *Uniform ambulatory medical care: Minimum data set* (DHHS Pub. No. (PHS) 81-1161). Hyattsville, MD: National Center for Health Statistics.

Ellwood, P.M. (1988). Shattuck lecture. Outcomes management: A technology of patient experience. *New England Journal of Medicine*, 318, 1549-1556.

Fetter, R., Youngoo, S., Freeman, J., and others. (1980). Case mix definition by diagnosis-related-groups. *Medical Care*, 19(2, Suppl.), 1-49.

Fetter, R.B., Averill, R.F., Liechtenstein, J.L., and Freeman, J.L.. (1984). Ambulatory visit groups: A framework for measuring productivity in ambulatory care. *Health Services Research,* 19, 415-437.

Fink, R. (1973). Analysis of utilization data. *Medical Care*, 11(Suppl. March-April), 109-116.

Fries, B. and Cooney, L. (1985). Resource utilization groups: A patient classification system for long-term care. *Medical Care*, 23, 110-122.

German, P.S., Skinner, E.A., Shapiro, S., and Salkever, D.S. (1976). Preventive and episodic health care of inner-city children. *Journal of Community Health,* 2, 92-106.

Gonnella, J.S., Hornbrook, M.C., and Louis, D.Z. (1984). Staging of disease: A case-mix measurement. *Journal of the American Medical Association*, 251, 637-644.

Graduate Medical Education National Advisory Committee (GMENAC). (1980). *Final report*, Vol. 1 (DHHS Pub. No. (HRA) 91-651). Hyattsville, MD: Health Resources Administration.

Horn, S.D. and Horn, R.A. (1986). The computerized severity index: A new tool for case mix management. *Journal of Medical Systems*, 10, 73.

Hornbrook, M.C., Hurtado, A.V., and Johnson, R.E. (1985). Health care episodes: Definition, measurement and use. *Medical Care Review*, 42(2), 163-218.

Iezzoni, L.I. and Moskowitz, M.A. (1988). A clinical assessment of MedisGroups. *Journal of the American Medical Association*, 261, 3159-3163.

Institute of Medicine. (1989). Gray, B.H. and Field, M.J. (Eds.). *Controlling costs and changing patient care: The role of utilization management.* Washington: National Academy Press.

Keeler, E.B., Buchanan, J.L., Rolph, J.E., and others. (1988). *The demand for episodes of medical treatment in the Health Insurance Experiment* (R-3454-HHS). Santa Monica, CA: RAND Corporation.

Keeler, E.B. and Rolph, J.E. (1982). *The demand for episodes of medical treatment: Interim results from the Health Insurance Experiment* (R-2829-HHS). Santa Monica, CA: RAND Corporation.

Keeler, E.B. and Rolph, J.E. (1988). The demand for episodes of treatment in the Health Insurance Experiment. *Journal of Health Economics*, 7(4), 301-422.

Keeler, E.B., Wells, K.B., Manning, W.G., and others. (1986). *The demand for episodes of mental health services* (R-3432-NIMH). Santa Monica, CA: RAND Corporation.

Kessler, L.G., Steinwachs, D.M., and Hankin, J.R. (1980). *Episodes of psychiatric utilization.* Medical Care, 18(12), 1219-1227.

Kessler, L.G., Steinwachs, D.M., and Hankin, J.R. (1982). Episodes of psychiatric care and medical utilization. *Medical Care*, 20, 1209-1221.

Kessner, D.M., Kalk, C.E., and Singer, J. (1973). Assessing health quality: The case for tracers. *New England Journal of Medicine*, 288, 189-194.

Knaus, W.A., Draper, E.A., Wagner, D.P., and Zimmerman, J.E. (1985). APACHE II: A severity of disease classification system. *Critical Care Medicine*, 13, 818-829.

Lasdon, G.S. and Sigmann, P. (1977). Evaluating cost-effectiveness using episodes of care. *Medical Care*, 15, 260-264.

Localio, A.R., Lawthers, A.G., Brennan, T.A., and others. (1991). Relation between malpractice claims and adverse events due to negligence: Results of the Harvard Medical Practice Study III. *New England Journal of Medicine*, 325(4), 245-251.

Lohr, K.N. and Brook, R.H. (1980). Quality of care in episodes of respiratory illness among Medicaid patients in New Mexico. *Annals of Internal Medicine*, 92, 99-106.

Moscovice, I. (1977). A method for analyzing resource use in ambulatory care settings. *Medical Care*, 15, 1024-1044.

Mushlin, A.I. and Appel, F.A. (1980). Testing an outcome-based quality assurance strategy in primary care. *Medical Care*, 18 (Suppl. 5), 1-100.

Roos, L.L., Cageorge, S.M., Austen, E., and Lohr, K.N. (1985). Using computers to identify complications after surgery. *American Journal of Public Health*, 75(11), 1288-1295.

Rupp, A., Steinwachs, D.M., and Salkever, D.S. (1984). The effect of hospital payment methods on the pattern of care and cost of mental health care. *Hospital and Community Psychiatry*, 35, 456-459.

Salkever, D.S., German, P.S., Shapiro, S., and others. (1976). Episodes of illness and access to care in the inner city: A comparison of HMO and non-HMO populations. *Health Services Research*, 11, 252-270.

Salkever, D.S., Skinner, E.A., Steinwachs, D.M., and Katz, H. (1982). Episode-based efficiency comparisons for physicians and nurse practitioners. *Medical Care*, 20, 143-153.

Scitovsky, A.A. and McCall, N. (1976). *Changes in the costs of treatment of selected illnesses, 1951-1964-1971.* NCHSR Research Digest Series (DHEW Publication No. (HRA) 77-3161). Rockville, MD: National Center for Health Services Research and Health Care Technology Assessment.

Shapiro, S., Steinwachs, D.M., Skinner, E.A., and Mushlin, A. (1976). *Survey of quality assurance and utilization review mechanisms in prepaid group practice plans and medical care foundations.* Baltimore, MD: The Johns Hopkins University.

Solon, J.A., Feeney, J.J., Jones, S.H., and others. (1967). Delineating episodes of medical care. *American Journal of Public Health*, 57, 401-408.

Solon, J.A., Rigg, R.D., Jones, S.H., and others. (1969). Episodes of medical care: Nursing students' use of medical services. *American Journal of Public Health*, 59(6), 936-946.

Starfield, B., Simborg, D., Horn, S., and Yourtee, S. (1976). Continuity and coordination in primary care: Their achievement and utility. *Medical Care*, 14, 625-636.

Starfield, B., Weiner, J., Mumford, L., and Steinwachs, D. (1991). Ambulatory care groups: A categorization of diagnoses for research and management. *Health Services Research*, 26(1), 53-74.

Steinwachs, D.M. (1979). Measuring provider continuity in ambulatory care: An assessment of alternative approaches. *Medical Care*, 17(6), 551-565.

Steinwachs, D.M. and Mushlin, A. (1978). The Johns Hopkins Ambulatory Care Coding Scheme. *Health Services Research*, 13, 36-49.

Steinwachs, D.M., Weiner, J.P., Shapiro, S., and others. (1986). A comparison of the requirements for primary care physicians in HMOs with projections made by GMENAC. *New England Journal of Medicine*, 314, 217-222.

Steinwachs, D.M., Weiner, J.P., and Shapiro, S. (1989). An expanding role for management information systems in quality assurance. In Goldfield, N. and Nash, D.B. (Eds.), *Providing quality of care: The challenge to the clinician.* Philadelphia: American College of Physicians.

Steinwachs, D.M. and Yaffe, R. (1974). *Development of patterns of ambulatory care from an encounter data information system.* Presented at the Operations Research Society Meetings, San Juan, Puerto Rico.

Steinwachs, D.M. and Yaffe, R. (1978). Assessing the timeliness of ambulatory medical care. *American Journal of Public Health,* 68(6), 547-556.

Stewart, A.L., Greenfield, S., Hays, R.D., and others. (1989). Functional status and well-being of patients with chronic conditions: Results from the Medical Outcomes Study. *Journal of the American Medical Association,* 262(7), 907-913.

Sytema, S., Giel, R., and ten Horn, G.H.M.M. (1989). Patterns of care in the field of mental health: Conceptual definition and research methods. *Acta Psychiatrica Scandinavica,* 79, 1-10.

Tarlov, A., Ware, J.E., Greenfield, S., and others. (1989). The Medical Outcomes Study: An application of methods for monitoring the results of medical care. *Journal of the American Medical Association,* 262, 925-930.

Weiner, J.P., Starfield, B.H., Steinwachs, D.M., and Mumford, L.M. (1991). The development of a population oriented case mix measure for application to ambulatory care. *Medical Care,* 29, 452-472.

Weiner, J.P., Steinwachs, D.M., Shapiro, S., and others. (1987). Assessing a methodology for physician requirement forecast: Replication of GMENAC's need-based model for the pediatric specialty. *Medical Care,* 25(5), 426-434.

Young, K.M. and Fisher, C.R. (1980). Medicare episodes of illness: A study of hospital, skilled nursing facility and home health agency care. *Health Care Financing Review,* 2(2), 1-23.

Young, W.W. (1984). Incorporating severity of illness and co-morbidity in case-mix measurement. *Health Care Financing Review,* 6(Suppl), 23-31.

Zawistowich, L. (1989). *Impact of changes in medicaid payment policies on psychiatric episodes of care.* Unpublished dissertation, The Johns Hopkins University.

Using Episodes of Care in Effectiveness Research

Emmett B. Keeler, Ph.D.

An episode is a theoretical construct that aggregates numerous smaller units of service. Researchers must assemble the data to be combined, write programs to partition the data into episodes, and create derived variables describing the episodes. In effectiveness research, as in other areas of health services research, the "episodes" approach has theoretical advantages but practical difficulties. Many theoretical advantages have been noted (Hornbrook, Hurtado, and Johnson, 1985; Steinwachs, in this volume), but researchers at RAND have encountered practical difficulties when using episodes of treatment in their work.

The conceptual advantages of episodes include a comprehensive accounting of inputs and a better understanding of the timing and alternatives of treatment decisions. Effectiveness research mimics clinical trials as closely as possible to enhance the credibility of its results. This approach is easier for disease treatments where onset is quick, everyone is immediately treated in the same location, and little followup is needed; but few conditions fit this mold. In reality, even surgical procedures have followup care, presenting patients vary, and not everyone comes to the hospital.

It is natural and attractive to relate all health care resources used in treatment to health outcomes. If a treatment plan combines resources from different settings, we have to evaluate that plan by aggregating all the inputs. Comparing the outcomes of varieties of surgery that are always inpatient may be easy, but when comparing the outcome of surgery with that of a longer course of medical care or to "watchful waiting," figuring out what goes into the treatment is more demanding.

Episodes mirror decisions. We may be able to get comparable cohorts of patients and date the start of "treatment" to reduce various biases that arise from starting just at an operation. Analyses that are based on units which reflect behavior will be clearer, stronger, and more credible, since the results will not be attenuated through improper aggregation. For issues of access, the patient's decision to get care may be paramount, and we need to have some way to measure that. For issues of treatment, it may be the initial visit, or shortly thereafter, when the diagnostic information comes in and a plan is made. If most patients with suspected disease get a certain test, it may be possible to look for that test as a starting point (for example, Dopplers for peripheral vascular disease) and even use the test information if it is available to adjust for severity.

With episodes, we can test the effects of quality or of alternate ways of providing the service. By looking at what happens within the episode, we can tell whether such differences matter. For example, in rehabilitation or mental health treatments, is quantity of care all that matters, or is there a difference in impact between weekly treatments for 20 weeks or daily treatment for 4 weeks? Often there are guidelines about timing of prenatal or postsurgical visits. Do these matter?

Embry Howell's (1991) study of prenatal care in France and the United States shows that France, with many fewer prenatal visits per mother but with nearly all visits starting in the first trimester, has better outcomes than the United States, even after adjustment for some other predicting factors. Looking simply at visits per mother would miss these results. Such results are important in deciding whether we want to increase outreach or the number of visits by those in the system.

The definition of episode can be tailored to the problem under analysis. We can study access by looking at relation of symptoms or at number or type of episodes and where they occur. We can study treatments by looking at what happens within episodes. An analysis of why blood pressure control was better for participants with free care in the Health Insurance Experiment (HIE), followed a chain: going to a doctor, getting a diagnosis of hypertension, having appropriate treatment prescribed, compliance, and control (Keeler, Brook, Goldberg, and others, 1985). The major difference was between hypertensives who never went to a physician

Dr. Keeler is a Senior Mathematician with the RAND Corporation, Santa Monica, CA.

over the course of the study and those with just a few visits and no diagnosis. Once hypertension was diagnosed, treatment and compliance were similar with free care and with cost-sharing.

We can study compliance by looking at timing and quantity of followup visits. In the HIE, we studied both the rate of initiation (access) and cost of episodes for economic analyses. For the quality of care analysis, we derived variables of the form, "If x is diagnosed, then in the next y weeks, was there workup (or followup or prescription) suitable for x?"

Episodes were discussed and used by Neu and Harrison (1986) in estimating the costs of posthospital Medicare covered care, for which they put together information on readmissions, skilled nursing facilities, home health, and office care. This work is now being extended to test access with different types of insurance. In the past, combining Part A and Part B data has been difficult, but the Health Care Financing Administration and private researchers have been working on this problem. An additional problem for those interested in the total costs of treatment—which is even more severe for younger patients—is that of missing data on services not covered by insurance from their primary source.

In the study of quality of care under prospective payment, we investigated whether quick discharge under a prospective payment system led to bad consequences in long-range institutionalization. Initial discharge to a nursing home was shown in the hospital medical record but was not necessarily bad and indeed was not strongly correlated with permanent institutionalization. Because we could not retrospectively determine from Part A data whether patients were in nursing homes 6 months later, we looked for a physician visit in Part B data during a window around 6 months after admission (Kahn, Keeler, Sherwood, and others, 1990).

Episodes of treatment are particularly attractive for the cost and economic side of cost-effectiveness. First, they are a "natural" unit of decision—medical care is sold in treatment packages—and for studying incentives that may change through the year, it is important to get the full cost and timing right. In general, behavioral models lead to stronger and more credible estimates of the effects of the variables influencing behavior, since the effects are not attenuated by improper aggregation.

Second, in modeling costs of care, it is convenient to model rates of treatment episodes and their costs separately and determine overall costs by multiplying rate by costs per episode. We found insurance affects only the rate, not cost, of treated episode. Episodes provide more detail than annual expenditures, allowing us to look within the year, but they offer some useful data reduction when compared with line charges or even visits. We found that acute and health maintenance episodes occur independently over time. This and the data reduction made statistical modeling and fitting much easier (Keeler and Rolph, 1988).

Practical Problems in Using Episodes

Steinwachs (in this volume) asks why if episodes are theoretically so valuable, they have not been used more in health services research. He suggests that a lack of standards and the difficulty of conceptualizing episodes for chronic conditions may be important. A third even more important reason is the time and effort needed to assemble a good episodes data base. Even in the HIE, a project with unusually relaxed deadlines and ample resources, we often wondered whether we would ever complete the episodes data file.

Assembling Data

In the HIE, we decided from the beginning that we would try to do an episodes analysis. We designed our claims forms to facilitate episode grouping. Each line charge was supposed to be linked to diagnosis and given a treatment history code (acute, chronic, chronic flareup, well, and for each of these four, initial or continuing). This system was very useful in separating out multiple simultaneous episodes. From one point of view, chronic flareups are part of the chronic condition, but from an economic point of view, they are unexpected. Although we had taken steps to prevent participants from using nonexperimental insurance, we were concerned about (noninsured) out-of-plan use. This is a general problem in assembling claims data, whether from health maintenance organizations (HMOs) or particular fee-for-service insurers. Because we had only a few sites for our patients, we were able to verify utilization by checking with the physicians' offices and hospitals in those sites for any records of our patients and estimate the amount of out-of-plan use.

In addition, for quality-of-care episodes, we could use information from health diaries, collected biweekly. These diaries provided information on untreated illness, telephone calls to providers, and reasons for visits. These data were also used to cross-validate claims data and estimate their underreporting. Even though participants were paid to send in these forms, such information did not appear to be as reliable as claims data.

Perhaps HMOs with little outside use can get such rich or complete information on use and patient health, but in general, it is a difficult task; creative effort is required to assemble such data on a set of patients.

Linking Data

The final economic episodes program was developed over a 5-year period with several years of full-time ef-

fort devoted to that task. The difficulty comes from the computer having to consider a lot of context in deciding where to place successive line charges and whether to start a new episode. Initial decisions about filling in missing treatment history codes and overwriting obviously incorrect codes were made by George Goldberg, M.D. He and Bonnie Scott, M.D., also advised us on when a gap represented a new episode and when it was probably a continuation. Researchers, programmers, and clinicians must work together to find appropriate compromises between what would be ideal and what the data and realistic time and costs of programming limits allow. After considerable analysis, two revisions of the program were useful for fixing subtle problems and deriving some other variables describing the episodes. (For the rationale and a description of the final set of programs that convert the line charges to a set of episodes, see Keesey, Keeler, and Fowler, 1985.)

Timing issues can be subtle. For example, in studying the effects of deductibles on behavior, we needed to determine the status of the family relative to their deductible at the time the decision was made. For a pregnant women whose obstetrical bills will exceed the deductible, the marginal cost of any additional care for pregnancy or indeed for any other medical condition is the postdeductible rate, thus influencing incentives from the time she knows she will deliver. Other timing issues involve careful thought about what patients know and when they know it. Such issues can be particularly difficult for long episodes, such as continuing chronic medical or mental health treatment.

Other difficult timing issues involve outcome measurement. Even for birth, where the start of the condition can be precisely fixed for all delivery "treatments," outcome measurement can be tricky. The outcome is presumably some aggregated function of health and satisfaction over time. How do we weigh the times before birth, during delivery, 2 days later, 1 month later, and 1 year later? How do we measure these outcomes? In comparing the outcome of arthroscopic surgery versus 6 months of rest, how do we combine outcomes that are 1 hour, 3 months, and 1 year past surgical treatment with the very different pattern of functioning and health of nonsurgical treatments? These problems are similar to well-known ones of mortality postadmission for regions with very different patterns of length of stay.

Creating Derived Variables

Because the new file will have episodes as its unit, the programs must define derived variables that describe the episodes for analysis. Some of these are fairly subtle, such as "intensity" of mental health services, "continuity of care," and "adequate followup." The definitions of such variables are the result of interactions between researchers, programmers, and data to see what is feasible and what degree of accuracy is acceptable. If the concepts are too complicated and difficult to program, it may be hard to know whether the program is working correctly.

Conclusions

Episodes are natural units of analysis for answering questions about the relation of health care inputs to outcomes. Once they have been developed, they make modeling easier and more plausible and have led to remarkable insights. Unfortunately, the effort needed to develop a good episodes data set for many situations is enormous; as in all research, compromises about data quality and costs must be made. Even in the HIE, the best funded health services research project we will see, we did not have the resources to get the line charges put exactly right into episodes. We lagged several years behind the simpler annual analyses of costs and could not include all the years completed by subjects in the field. Nevertheless, the episodes data led to insights not attainable from the annual analyses about how people made decisions and especially how they reacted to deductibles.

Because of differences in data and in analytic purpose, it is difficult to arrive at uniform standards and definitions. Additional information can be very useful in linking episodes. Since data may differ in their suitability and the task is so hard, we have to use what we have to the limit.

Although we had the same data and largely the same researchers working on utilization and quality of care in the HIE, we were unable to come up with episode data that could be used for both analyses. The studies had different episode concepts and hence two different elaborate programs. In each case, the building block was the billing line, but they were put together in very distinct ways. A third and simpler program was developed for the economic analysis of outpatient mental health care (Keeler Manning, and Wells, 1988).

Outpatient mental health episodes were much easier to program, since people typically went for care on a regular basis to the same provider, and there was only one line charge to a visit. The economic episodes differed from quality-of-care episodes because of the economic issues that relate to accounting year effects and affect economic decisions but do not affect disease pathology. (Indeed, in the mental health episodes, we noted a higher than normal drop in use of services at the end of the year for those who had to meet another deductible before their mental health care was free.) From an epidemiologic point of view, the accounting year was a nuisance, but it did have some impact on decisions. Differences in

settings may also cause difficulties in uniform definitions: in the HMO studied by Kessler (1980), mental health visits were very rarely less than 1 week apart (perhaps due to an HMO policy), whereas in fee-for-service, smaller breaks were fairly common. The distribution of breaks observed has considerable effect on false positive rates of new episodes if these are defined by breaks of a certain length.

References

Hornbrook, M.C., Hurtado, A.V., and Johnson, R.E. (1985). Health care episodes: definition, measurement and use. *Medical Care Review*, 42, 163-218.

Howell, E.M. (1991). Socioeconomic factors in birth outcomes: The United States and France. Presented at the annual meeting of the Association for Health Services Research, San Diego, CA, June 1991.

Kahn, K.L., Keeler, E.B., Sherwood, M.J., and others. (1990). Comparing outcomes of care before and after implementation of the DRG-based prospective payment system. *Journal of the American Medical Association*, 264, 1984-1988.

Keeler, E.B., Brook, R.H., Goldberg, F.A., and others. (1985). How free care reduced hypertension in the Health Insurance Experiment. *Journal of the American Medical Association*, 254, 1926-1931.

Keeler, E.B. and Rolph, J.E. (1988). The demand for episodes of treatment in the health insurance experiment. *Journal of Health Economics*, 7, 337-367.

Keeler, E.B., Manning, W.G., and Wells, K.B. (1988). The demand for episodes of mental health services. *Journal of Health Economics*, 7, 369-392.

Keesey, J., Keeler, E.B., and Fowler, W. (1985). The episodes of illness processing system (RAND N-1745-1-HHS). Santa Monica, CA: RAND Corporation.

Kessler, L.G., Steinwachs, D.M., and Hankin, J.R. (1980). Episodes of psychiatric utilization. *Medical Care*, 8, 1219-1227.

Neu, C.R. and Harrison, S. (1986). Prospective payment for medicare posthospital services: Some empirical considerations (RAND R-3435-HCFA). Santa Monica, CA: RAND Corporation.

Medical Effectiveness Research Data Methods: Summary and Reactions

Joseph P. Newhouse, Ph.D.

Introduction: Reliability and Validity of Existing Administrative Data

Many of the papers in this volume raise issues about the reliability and validity of administrative data, including those of Iezzoni, Romano and Luft, Steinwachs, and Turner, Markson, Houchens, and Fanning. All of these authors give a comprehensive view of the issues. Among the problems they emphasize are:

1. The issue of errors in diagnosis and procedure codes, some of which occur for financial reasons.[1] A dramatic example is the coding of ventilator support when diagnosis-related group (DRG) 475 was established in 1988 to pay more for cases with ventilator support; the coding of ventilator support increased by an order of magnitude (Carter, Newhouse, and Relles, 1990). The presumption is that ventilator support was previously undercoded.[2]

2. Ambiguities in the coding classification, as emphasized by Iezzoni and Romano and Luft. Such ambiguities, of course, are not limited to the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9-CM) system but are also found in the Health Care Financing Administration (HCFA) Common Procedure Coding System (HCPCS) (or CPT-4) coding for physician services, as witnessed by the lack of standardized meaning for office visit codes (Braun, Hsiao, Becker, and others, 1988).

3. Limitation on the number of diagnoses and procedures (see Romano and Luft) that are recorded on the face sheet. This latter limitation can create severe difficulties in adjusting for comorbidities.

Additionally, Iezzoni makes the point that reporting in the chart may become "increasingly gloomy" by the greater need to justify utilization. It is difficult, however, to know how much this reflects bias and how much it reflects the true state of the world. Not only may there have been underreporting in the chart previously, but we know also that Medicare patients were sicker at admission in the 1985-1986 period than in the 1981-1982 period (Keeler, Kahn, Draper, and others, 1990).

There are, of course, issues of reliability and validity for data other than diagnosis and procedure. A few of these include:

1. If utilization occurs, are claims filed? In the Medicare Part B data, reporting is reasonably complete for relatively well-reimbursed procedures (Kosecoff, Chassin, Fink, and others, 1987), and one would expect the Part A data to be at least as complete. I do not know of similar studies for less well-reimbursed procedures, although health maintenance organization (HMO) data are known to be incomplete, as Romano and Luft emphasize.

   Medicaid claims data, however, may well be less complete; certainly there are numerous anecdotes about Medicaid rates being too low to justify the physician filing a claim. As Medicaid files become more accessible over the next several years, a high priority for outcomes research is to establish the completeness of reporting. Because completeness is

Dr. Newhouse is John D. MacArthur Professor of Health Policy and Management and Director of the Division of Health Policy Research and Education at Harvard University.

[1] Iezzone and Romano and Luft (in this volume) point out errors because of the incentives for upcoding, citing the study by Hsia, Krushat, Fagan, and others (1988). This study used data from 1985, but data from subsequent years show a marked reduction in bias after 1985 (that is, errors favor the Government approximately as often as they favor the hospital) and a gradual improvement in agreement between the hospital coding and the coding of a disinterested coder (Carter, Newhouse, and Relles, 1990).

[2] Contrary to Turner, Markson, Houchens, and Fanning, a more complete medical record or more complete coding of the record, as well as shifts in the order of conditions, may cause "DRG creep."

likely to vary with the level of reimbursement (for example, cesarean sections are more likely to be reported than injections), one would like to estimate underfiling as a function of reimbursement. The extent of fraud may also be an issue, but one that appears difficult to research.

2. A different sort of issue arises with pharmaceuticals—did the beneficiary use them or resell them? There are accounts that when New York's Medicaid program required triplicate filing for certain drugs, prescriptions of diazepam fell 85 percent for Medicaid beneficiaries in the first 6 months versus a 20-percent fall statewide (David Axelrod, personal communication, August 1990). It is difficult to explain this discrepancy unless the drugs were being resold.

3. In many American administrative data sets, it is difficult to obtain data on families. These data include both family composition (of importance, for example, in studies in which a decision to institutionalize is at issue) and family level variables such as income or parent's education when the utilization record is for the child (see Zambrana, in this volume, for example). Additionally, for the purpose of defining a poor population, one would like a longer term measure of income than the current year's income.

4. Ethnicity may not be reliably coded in many administrative data sets (again, see Zambrana in this volume).

Additional Data

Having mentioned some difficulties with the administrative data that have historically been one backbone of this field, I now want to turn to the issue of how such data might be improved for the purposes of outcomes research. A main point is that major improvements can be made in outcomes research if additional data are available.

How data availability influences analysis is emphasized by contrasting the analyses of Canadian data discussed in the paper by Roos, Wajda, Nicol, and Roberts (in this volume) with the analyses discussed in the other papers reported here that focus on American data. Although the Canadian data are not perfect, they have the following advantages over American data:

1. Most importantly, there is a longitudinal record of utilization in the under-65 population. The United States, of course, has somewhat similar data for those 65 and over because of the near universality of Medicare, but: some procedures—most notably reproductive procedures—are not found in the elderly or are found in such small numbers as not to be useful; and results found in the elderly may or may not generalize to the under-65 population.

2. The Canadian coverage is more complete—in particular, drugs used by outpatients are covered, whereas they are not covered by Medicare. This obviously creates a problem for using Medicare data to study the effects of pharmaceuticals in the United States, as well as the outcome of any procedure in which the outcome depends on which drugs were given when.

3. A complete (100 percent) sample of data is available for outpatient services. Only a 5-percent sample of the Medicare data is readily available on a national basis.

Despite these advantages of the Canadian data, some data are not readily available from either source, leading to the issue of which data we want to have. Three improvements seem feasible.

One improvement would be to greatly augment current abilities to obtain a nationwide random sample of patient charts. The place to start in my view is with inpatient charts of Medicare patients, because a small random sample already exists that could be increased.[3] Because a population sample of charts may well not have sufficient numbers to study a particular disease or treatment, there should be the capability to enhance the sample with particular patient types.

A sample of charts would serve several purposes for outcomes research:

1. It would provide data on clinical values of tests, thereby allowing better standardization across patients. This is not a unanimous view. For example, Roos, Wajda, Nicol, and Roberts (in this volume) assert that "claims can often generate case-mix adjustments almost as good as or as good as those provided by expensive primary data collection." Nonetheless, this has not been the prevailing view in the United States and, as Iezzoni (in this volume) emphasizes, is certainly not the case for such items as physical or psychological functioning or socioeconomic factors (the chart may also be uninformative about the latter). The assertion is also belied by the emphasis Selker, Griffith, Beshansky, and others (in this volume) put on electrocardiogram (ECG) and left ventricular ejection fraction data in their predictive model for patients with acute myocardial infarction (AMI).

2. The sample would provide data on services that are not covered—most importantly drugs—as well as

[3]This is a sample at the SuperPRO organization.

some procedures where the reimbursement may be regarded as too trivial to merit billing. (In an outpatient setting, however, patient compliance with a prescribed drug would remain unknown.)

Other purposes would also be served by a sample of charts that relate to operational concerns of insurance programs. For instance, in Medicare, one could estimate how much of the annual change in the Case Mix Index is upcoding. One could also better estimate the quality of care being delivered by various institutions—for example, HCFA's risk-adjusted mortality indices. These purposes strengthen the case for a sample of charts.

A second potential improvement is to add information to the claims forms that would permit better definition of episodes, the advantages of which are well stated by Steinwachs (in this volume). Several additional data items would permit better classification. One could ask for multiple diagnoses on outpatient claims forms and ask the physician to link any procedures or tests to the listed diagnoses (this would resolve ambiguity about the process of care for patients with multiple diagnoses). One could also ask for treatment history codes; for example, is this visit a routine visit for a chronic condition or for a flareup of a chronic condition? Further, one could ask for the reporting of any drugs or supplies ordered and that they be linked to diagnoses.[4] This would be a cheaper way to obtain information about drugs, especially outpatient drugs, than retrieving charts. Coding such information would be relatively expensive, but one could begin by not coding the information routinely but only for special studies.

A third improvement would be to move to a 100-percent sample of Part B data at the national level, obviating the need for each research team to deal with each carrier if the 5-percent Part B Medicare annual data (BMAD) beneficiary file does not contain enough of the desired cases.

These are all practical steps that would both improve the management and operation of insurance programs and enhance outcomes research. Various authors (in this volume) discuss a number of other problems posed by administrative data that these steps would not resolve. Two in particular are worth mentioning:

1. These steps would obviously not resolve the difficulties in obtaining administrative data on the under 65 on a population basis. This, however, is a byproduct of our lack of universal insurance arrangements, there being no analog of claims data on the uninsured. Thus, fully resolving this issue will not happen in lieu of universal insurance. For some purposes, data collected by hospital discharge services or claims data from large employers may be a substitute, but the usefulness of these data must be appraised on a case-by-case basis. In particular, variation across selected hospitals is more likely to reflect unmeasured case-mix differences than variation across large areas (such as States). Claims data from large employers often lack information on the population eligible for coverage, and individuals change employers, thereby inhibiting followup. Medicaid data sometimes may be a substitute, but in addition to the potential underfiling problem already noted, the lack of continuous eligibility makes their use more problematic than the use of Medicare data. Hospital discharge data sets can be a substitute if everything of interest is contained within the hospital stay. If these data do not suffice, I do not see any inexpensive answer to this issue.
2. The three steps outlined above will not solve the issue of obtaining measures that come from the patient, including patient satisfaction measures and measures of morbidity such as functional status. Those appear to be obtainable only from primary data collection. So too are many data on sociodemographic status, the data from claims being limited to such surrogates as self-pay or Medicaid status.

The Issue of Cost

An issue given little emphasis in all the papers reported here, but likely to become increasingly important, is the accuracy of data on cost and price. The three steps described above will also not resolve this issue. To some degree, the issue of cost could be considered a subset of the questions related to the validity of administrative data, but because the papers do not raise it in that context, I am singling it out for special emphasis.

Clinicians often accord less importance to costs than economists, so let me briefly note why costs are important.[5] The reasons fundamentally relate to getting the most benefit from a given commitment of resources. From a research perspective, this is the basis of cost-effectiveness and cost-benefit analysis. From an operational perspective, prices in medical care tend to be administratively set (namely, Prospective Payment System [PPS], Resource Based Relative Value Scale [RBRVS], and Adjusted Average per Capita Cost [AAPCC]) and to the degree these prices deviate from cost, we send false signals to the medical care sector

[4] These pieces of information were on the claims form used in the RAND Health Insurance Experiment and were important in constructing episodes, allocating costs to episodes (see below), and judging quality of care (Keeler and Rolph, 1988).

[5] For example, the RAND studies on appropriateness do not take explicit account of cost, but rather they use as a definition that the benefit to the patient should outweigh the risk of the procedure.

about how to produce medical care, just as subsidies of various agricultural inputs send false signals to farmers and can cause important inefficiencies in agriculture. Indeed, reducing such distortions is at the bottom of the rationale for the RBRVS reforms of physician payment. It is therefore important to emphasize that existing measures of cost in general relate not to the conceptually desirable measure of cost—which is what society gives up in foregone opportunities when a particular service is or is not produced—but rather to an accounting definition of cost or to charges, both of which may be seriously misleading.

Charges are a list price, but almost no one pays hospitals at list price. Perhaps as a result, there are well-known cross-subsidies among various hospital departments (Harris, 1979). Because each disease does not draw on the services of the various departments equally, distortions can arise in costing the treatment of various patients. For example, patients who have had an acute myocardial infarction (AMI) use intensive care unit days relatively more than the average patient. Such services are marked up less than other ancillary services, such as pharmacy, laboratory, and radiology. Thus, use of charges for AMI patients is likely to understate their costs relative to other patients. Moreover, there is no assurance that these biases are similar across institutions.

Some of the problems with cost include:

1. Usually, the cost figures used are operating costs and exclude capital costs (or if capital costs are included, they are included at historical cost, which may be highly understated). Capital costs will not be uniform by type of problem or disease; for example, cancer patients receiving radiotherapy are likely to have above-average capital costs.

2. Costs are typically estimated by applying a departmental cost/charge ratio to charges, but the cost to charge ratio may vary substantially by procedure or service within a department. The resulting cross-subsidy problem is analytically the same as that described for charges. Presumably, matters are even less accurate if a hospital rather than a departmental cost/charge ratio is used.

3. Other types of cross-subsidy problems are caused by costing all hospital days in regular units at one rate (and analogous problems come from costing days in special units at a different rate). The last day of the stay uses less nursing time than the early days, which means the true saving from reducing length of stay is less than the average daily cost, which is presumably the cost that is used. Costs attributable to specific diseases may be distorted. AMI patients, for example, may use more nursing time than patients with pneumonia, but this difference is not reflected by using the same daily rate for both. Another example of the same problem is a uniform facility charge for the emergency room; with such an arrangement, it is likely that nonemergency cases subsidize emergency cases. Thus, if we shift nonemergency cases out of the emergency room, we will not save as much as the charge data tell us; put another way, the facility charge for the cases that remain is likely to rise.

4. Because one is usually considering changes at the margin (that is, somewhat more or somewhat fewer procedures) rather than eliminating a procedure altogether, the proper measure of cost is marginal cost, which may or may not bear much relationship to average cost, the figure usually used (for example, in costing the reduction in length of stay, as discussed above). Analogously, in the context of multiple products, the proper measure of cost is usually the incremental cost of that product, which means one does not want to allocate joint costs to specific services. Examples of costs that might be improperly allocated include the administrator's salary or the physician's telephone bill, both of which are likely to continue at the same level whether or not the specific service is delivered.

5. How costly it is to produce one service or procedure may depend on what other services or procedures are produced by the hospital or physician (or nursing home or home health agency).[6] Most current cost studies simply average in such economies of scope.

6. It is problematic to use the existing DRG weights as a measure of relative cost, both because of the difficulties just described and because many groupings are not homogeneous (for example, have coefficients of variation over 1). Moreover, weights for the under 65 have not, to my knowledge, been calculated on a national sample comparable to the Medicare sample.

Obtaining accurate data on marginal economic cost is not only likely to be difficult but also may need to be updated often. In light of their importance to operating administered price schemes such as the PPS, however, cost data have importance far beyond their usefulness in outcomes research.

The Issue of Area Definition

Most outcome studies have compared individuals in various geopolitical jurisdictions. The original studies began with small jurisdictions (Wennberg and Gittelsohn, 1973). Starting with the 1986 study by Chassin, Brook, Park, and others, however, it was shown that there were substantial variations in usage

[6]Economists call these economies or diseconomies of scope.

among large geographic areas. This finding was important, in that it became much less likely that the variations observed were attributable to some unobserved characteristics of patients or simply to chance.

Almost no one, however, seems to have given much thought to the optimal size area to use as an observation in outcomes research. One wants large enough areas so it is plausible that the distribution of unobserved variables is similar; smaller areas, however, will minimize within-group variance and thus enhance power.

The problem is more difficult because the variation we observe is likely provider-related. We do not, however, want to observe patients grouped by provider because they may differ in unobserved characteristics (some physicians, for example, may have more patients with complicated problems). Hence, observations are typically on patients grouped by area of residence. These patients may concentrate their usage among a set of providers to a greater or lesser degree depending on the size of the area used as a unit of observation.

How serious this issue is will depend on the true size of the effect we are seeking to measure. The optimal size area will almost certainly vary according to the problem being studied. For the moment, it is probably best to show that results are or are not sensitive to the area chosen as the unit of observation. Given the successes of outcomes research to date, it is not a problem that must be solved before progress can be made, but addressing it is likely to enhance outcomes research.

References

Braun, P., Hsiao, W.C., Becker, E.R., and others. (1988). Evaluation and management services in the resource-based relative value scale. *Journal of the American Medical Association*, 260, 2409-2417.

Carter, G.M., Newhouse, J.P., and Relles, D.A. (1990). How much change in the Case Mix Index is DRG creep? *Journal of Health Economics*, 9, 411-428.

Chassin, M.R., Brook, R.H., Park, R.E., and others. (1986). Variations in the use of medical and surgical services by the Medicare population. *New England Journal of Medicine*, 314, 285-290.

Harris, J.E. (1979). Pricing rules for hospitals. *Bell Journal of Economics*, 10, 224-243.

Hsia, D., Krushat, W.M., Fagan, A.B., and others. (1988). Accuracy of diagnostic coding for Medicare patients under the Prospective Payment System. *New England Journal of Medicine*, 318, 352-355.

Keeler, E.B., Kahn, K.L., Draper, D., and others. (1990). Changes in sickness at admission following the introduction of the Prospective Payment System. *Journal of the American Medical Association*, 264, 1962-1968.

Keeler, E.B. and Rolph, J.E. (1988). The demand for episodes of treatment in the health insurance experiment. *Journal of Health Economics*, 7, 337-367.

Kosecoff, J., Chassin, M.R., Fink, A., and others. (1987). Obtaining clinical data on the appropriateness of medical care in community practice. *Journal of the American Medical Association*, 258, 2538-2542.

Wennberg, J.E. and Gittelsohn, A. (1973). Small area variations in health care delivery. *Science*, 182, 1102-1108.

Learning from the Experts: What We Know About Effectiveness Research

Barbara Starfield, M.D., M.P.H.

Medical effectiveness research is not new. What is new is the concerted effort to develop such research into a defined field of endeavor, with goals and guidelines. In the few years since its definition as a field of study, medical effectiveness research is ready to take its place along with the controlled clinical trial as a major contribution to the armamentarium of tools for scientific inquiry.

The papers in this volume are testimony to the extraordinarily high quality work now contributing to knowledge about what works in medical practice and what does not and about what works better. Credit must be given to the pioneers who created the framework from which it derives, to the tenacity with which they pursued their ideas, and to the responsiveness of the Agency for Health Care Policy and Research (AHCPR) in transforming a grand design into a reality. The authors of the papers in this volume should be very proud of the contributions they have made by synthesizing the accomplishments of medical effectiveness research, its pitfalls, and its potentials. They have provided us with a good basis from which to move forward, and they deserve hearty congratulations.

Medical effectiveness research, or MEDTEP for short, has as its goal the improvement of medical practice "by developing and disseminating scientific information regarding the effects of presently used health care services and procedures on patients survival, health status, functional capacity, and quality of life" (AHCPR, 1990). It must involve a literature review and synthesis, use existing data to analyze variations in medical practice and associated patient outcomes, disseminate findings about effectiveness, and evaluate the effects of dissemination, particularly on voluntary behavior of providers. The outcome must be measured in terms of "reduced variation in practice patterns, more appropriate use of health care resources, and improvements in patient outcome." In order to accomplish these missions, the research team must consist of individuals who contribute a variety of skills including clinical expertise, epidemiology, biostatistics, research design, economics, decision analysis, survey research, data management, research synthesis, and research dissemination. Nothing in the general guidelines for medical effectiveness research requires multi-institutional collaboration; yet it is apparent—both from the requirements for broad expertise and from the broad charge to researchers—that such collaboration is necessary.

The first lesson we have learned from MEDTEP is that multi-institutional collaboration may be part of the secret of its success. The extent of multi-institutional collaboration is staggering. Selker and his coworkers (in this volume) describe the use of data bases from several institutions under conditions where data from a variety of separate trials have to be standardized and made comparable. This collaboration has a host of benefits. It makes investigators much more aware of ongoing work in the field and much more appreciative of its strengths and limitations. It provides much more support and expertise that can be applied to the research problem. It provides much larger sample sizes and more diverse populations for study. One of the contributions of MEDTEP has been the enhanced opportunity to take data from controlled clinical trials of efficacy supported by the National Institutes of Health (NIH) and apply them to a broader purpose: the study of effectiveness. This alone, however, would not do much to overcome a main limitation of controlled clinical trials: their essential nonrepresentativeness.

Romano and Luft (in this volume) mentioned but did not elaborate on the nonrepresentativeness of patients in controlled clinical trials. It stands to reason that when people volunteer to participate in research, whether before or after randomization, there is a selection bias operating. This selection bias, although not well studied,

Dr. Starfield is Professor and Head, Division of Health Policy, School of Hygiene and Public Health, The Johns Hopkins University.

may well operate in very critical areas, including comorbidity. More or less ill people may be more or less likely to participate in research; the direction of the effect is unknown and may not even be consistent across types of trials. When research stems from voluntary participation of institutions, both the patient and the practitioner populations are selected. Controlled clinical trials involving several institutions go a long way toward reducing this bias, but they can never get around the fact that the participating institutions are volunteering their participation. Medical effectiveness research has taken the quest for generalizability one step further in its use of existing population-based data. In medical effectiveness research using these data bases, there is no selection by institution. Everyone participates, albeit in ways that protect their privacy and confidentiality. Despite its limitation in ability to explain effects, medical effectiveness research has something that will never be attainable through controlled clinical trials: generalizability. This is one of the unique potential contributions of health services research, and it has been well realized in many MEDTEP studies.

Several of the papers in this volume make reference to the need to enhance existing data sets with clinical data—the second lesson we are learning from MEDTEP. The need for clinical data arises primarily because of the difficulty in obtaining information on most risk factors that are critical covariates in analytic research. The challenge is more than just inaccessibility of clinical data; it is also the unavailability of methods to characterize risk. Iezzoni's (in this volume) brilliant synthesis makes it clear that the appropriateness of a method of specifying risk depends on the purpose for which it is intended, but comorbidity is one risk factor that is always relevant. Comorbidity has generally been considered to mean diagnoses that are unrelated in etiology to the principal diagnosis. Most existing methods of assessing comorbidity, which Iezzoni summarizes, have been developed for use in hospitalized patients for purposes related to adjusting for the impact of these diagnoses on responsiveness to treatment of the principal diagnosis. Comorbidity probably has a broader impact, however. Recent studies (Starfield, in press; Starfield, Katz, Gabriel, and others, 1984) show that illness seems to predispose to other types of illness. It is not only that having one condition modifies the response to treatment of another; it is either that different illnesses have common causes or that illness begets illness (perhaps by altering resistance), even illnesses of other types. So comorbidity is an even more important risk factor than it might seem when just conceived as a covariate in treatment effectiveness.

Iezzoni goes beyond the biomedical concepts of acute clinical stability, severity of principal diagnosis, and extent and severity of comorbidities to define other factors that have to be considered in risk adjustment. These include physical functional status; psychological, cognitive, and emotional functioning; and sociodemographic attributes that are associated with illness levels. It is quite clear that biomedical characteristics are not the sole determinants of risk; it is not even clear that they are the main determinants.

Broadening the concept of risk to include these nonmedical factors presents even more of a challenge than does comorbidity. Limitations on access to medical records is surely one barrier; but even if medical records could be accessed, such information is unlikely to be found there, and when it is, it will be in unstandardized form which precludes its use in research. Broadening the concept of risk also will entail major modifications in recordkeeping. There clearly needs to be a concerted effort to enhance the content of medical records with some minimum core of generic information concerning, at least, major sociodemographic risks, including those related to social class, physical capacity, psychological well-being, general perceptions of level of health, and role functioning. Many individual clinical facilities already require providers to record some of these attributes. Here is an area where professional societies could take the lead. It would not be difficult to demonstrate to insurers and managers that they would gain by having such information available if it were feasible to do so.

Requirements for information on health status may not be long in coming. The Omnibus Budget Reconciliation Act of 1989 directed States to produce annual information on the health of children and youth within the State. State of health is defined more broadly than just the presence of particular illnesses. Some particular elements were specified, among them the number of children with chronic illnesses and their type, proportion of infants born with drug dependency, and proportion of 2-year-olds who are not adequately immunized. None of these items of information is readily available (Starfield, in press). If States are to carry out their mandates, they will have to devise some definitions (for example, what is a chronic illness?), some way of assessing their existence, and a way of obtaining the information, since there is no existing repository for these types of data. Perhaps it is no accident that about 20 States are already considering legislation to develop a single-payer system of insurance. The reasons for doing so are many, but among them must be the current frustration with the inability to account for dollars spent by documenting changes in health status in the population.

The third lesson we are learning from MEDTEP is the importance of considering the outpatient setting as well as the hospitalization. Several of the papers in this volume allude to the importance of outpatient data; the issue is central to Steinwachs' (in this volume) insightful assessment of the potential of episode analysis. Conventionally, episode analysis has been based on episodes of illness as defined by the presence of a diagnosis. Steinwachs' paper makes abundantly clear the limitations of such a conceptualization. Effectiveness analysis must ultimately start with the patient's problem, not with the diagnosis. One of the original patient outcomes research teams (PORTs)—the one concerning back pain—did so. Some second generation PORTs—the one on diabetes and the one on pneumonia—focus on diagnoses but within a context that maximizes the likelihood that all patients with the disorder are identified; the projects are outpatient-oriented. Since the outcomes of interest in PORTs are much broader than biomedical responsiveness, it is illogical that the process of analysis begin with the diagnosis of a problem defined in biomedical terms.

The figure specifies the important variables in health services research. It characterizes the health services system in ways that identify the major components of structure, process, and outcome. In the middle of the diagram is the process of care, which is made up of elements contributed by providers and those contributed by patients. The contribution of patients to the process of care starts with utilization; the contribution of providers starts with problem recognition. When Steinwachs (in this volume) speaks of the "episode," he speaks to the need for starting the process of care where it should be started, at the stage of problem recognition. As the capacity of MEDTEP moves forward, it should move the concept of the process of care back to problem recognition.

There must be a mine of information in studies of variation in physicians' recognition of people's problems. Rich lodes can already be found in the literature on physician-patient communication and practitioners' recognition of problems. Evidence is accumulating on the impact on outcomes of physician-patient agreement on the nature of the patient's problems (Kaplan, Greenfield, and Ware, 1989; Starfield, Steinwachs, Morris, and others, 1979; Starfield, Wray, Hess, and others, 1981), but the work has not yet been applied in a medical effectiveness framework.

An outpatient orientation with a focus on problems rather than diagnoses will go a long way toward dealing with the systematic undervaluation of the problems of the disenfranchised and minorities that Zambrana (in this volume) identifies in her paper. Only a small proportion of the population is hospitalized in a defined period of time, and not everyone is hospitalized in a lifetime. Almost everyone, however, has contact with some part of the health system. In 1 year, about 75 percent of the population consults a physician; in 2 years, almost 90 percent do so. Within a period of 5 years, virtually everyone has had some relationship with some practitioner at least peripherally involved in the formal or informal health network. If all parts of the health system kept a minimum set of information about patients, their problems, and their diagnoses, there would be a rudimentary data base of problems suitable for epidemiologic and health services research on whole populations; this data base would include virtually everyone.

Where they exist, outpatient data bases already have enough information to characterize populations by their diagnoses just using the techniques now available. Three types of approaches may be useful in categorizing case mix for purposes of MEDTEP research:

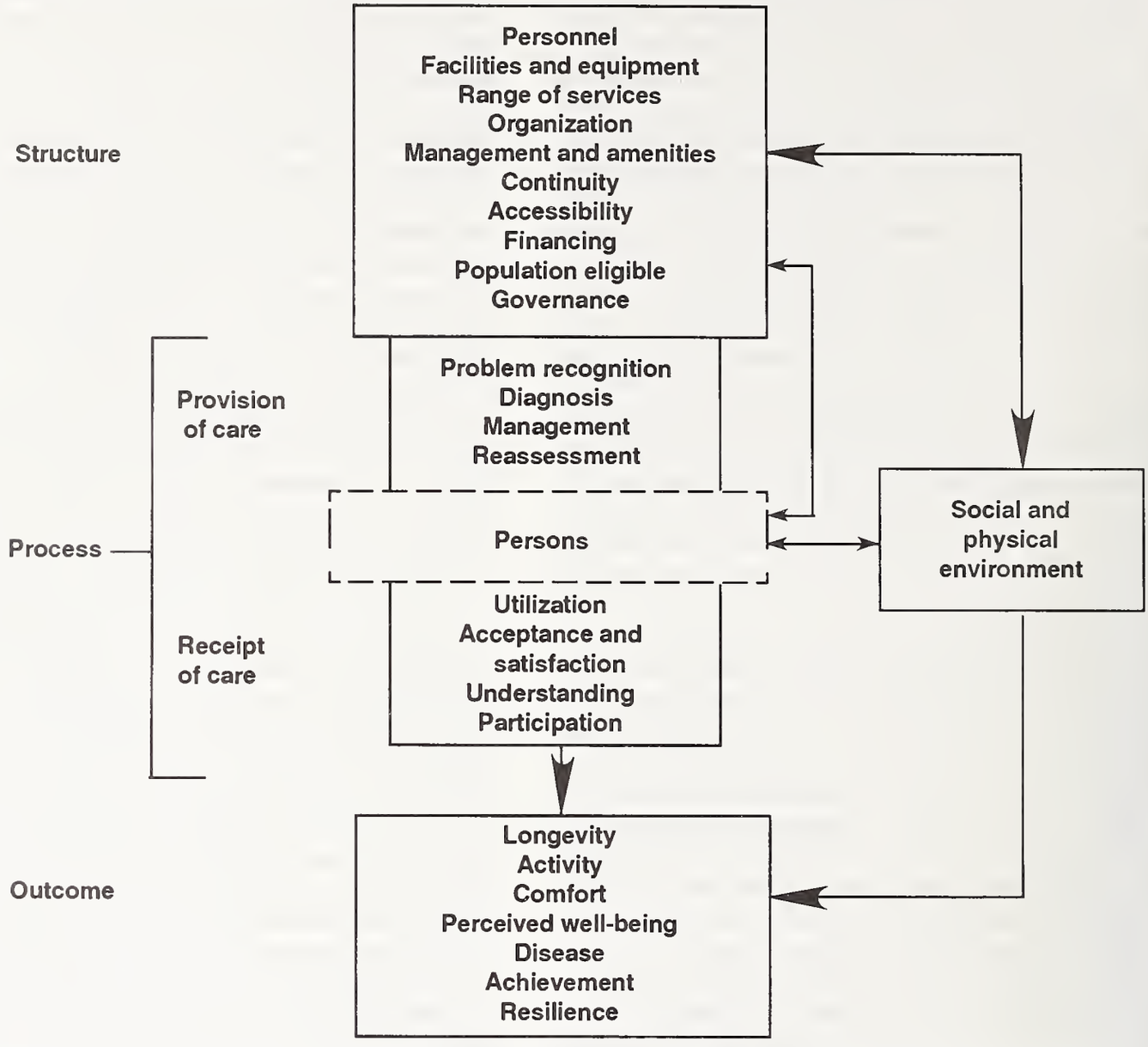
1. Characterize case mix at a particular point in time;
2. Categorizes it at any given point in time; and
3. Characterize case mix over a period of time.

Perhaps the best known of the first type is the staging-of-disease method. Staging has been useful in characterizing the clinical severity of diagnoses at a particular point in time: admission to the hospital. Examples of the second and third types are not so well known.

Second is the point-specific measure, usually used to categorize "case mix" from ambulatory visits. Investigators at Yale developed the Ambulatory Visit Group (AVG) system based on data from the National Ambulatory Medical Care Survey, which originally involved the use of duration of visits as the proxy measure of intensity or severity. This system is being refined through work supported by the Health Care Financing Administration. Because the method is designed for hospital outpatient visits, it is heavily oriented towards technologic procedures. In contrast to the original AVGs and most other case-mix systems, any given patient may have more than one APG for a given visit (Averill, Goldfield, McQuire, and others, 1991). The Products of Ambulatory Care system (Tenan, Fillmore, Caress, and others, 1988) used in New York State is a system used primarily for payment purposes; it does not contain much clinically relevant data.

In contrast, the Ambulatory Severity Index (ASI) is a method that captures information from both the physician and the medical record. The physician indicates the primary diagnosis, other diagnoses addressed in the visit, and other diagnoses not addressed in the visit. It captures the main reason for the visit and provides an assessment of complexity factors involving the patient's

Figure. The health services system



Source: Starfield, B. (1992). Primary care: Concept, evaluation, and policy. New York: Oxford University Press. Used with permission.

psychological and cognitive characteristics, physical limitations, clinical urgency, and complications of medical care. It has been tested in a number of settings—including a general medical clinic, an emergency medicine facility, and pediatric office practices—and seems to capture the essential differences in clinical status that would be expected in diverse settings.[1]

The Ambulatory Care Group (ACG) method is an example of the third type, which characterizes a pattern of diagnoses over a period of time. All codes in the International Classification of Diseases, Clinical Modification (ICD-CM) have been assigned to 1 of 34 different types, initially on their likelihood of persistence but then subdivided according to their likely severity and need for specialist versus generalist care. Computer algorithms employed data from several large health maintenance organizations and one State Medicaid plan to arrange and combine these diagnoses, with age and gender where relevant, into 51 unique combinations that characterize the mix of types of morbidity experienced by each individual over a year. It should be noted that one category of morbidity is signs and symptoms, so that "problems" as well as diagnoses are part of the system, depending of course on the existence of appropriate codes and the practitioner's diligence in recording the

[1]Horn, S., Sharkey, P., Starfield, B., and others. (In preparation). Differences in severity of illness across ambulatory settings.

problems so they can be coded. So far, the method has been applied to predictions of utilization of ambulatory services and to ambulatory and total costs for services over the current year and subsequent year. It has also been applied in epidemiologic analyses of burdens of morbidity experienced by individuals in different populations.

Any good risk assessment methodology needs to consider ambulatory care characteristics, even when applied to study hospitalizations. Therefore, point-specific risks have to be complemented by time-specific risks. The combination of various aspects of some of the different ambulatory case-mix methods deserves exploration. For example, the ASI—a point-specific method—could be complemented by the ACG (a time-specific method). There are a host of exciting possibilities for creative researchers to work together on these methodological issues.

Another potential contribution to clinical status assessment in the outpatient setting may accrue as a result of changes in the ICD. ICD-10 (10th Revision) will have a new format, with chapters. Chapter XXI in ICD-10 is entitled "Factors influencing health status and contact with health services." Its "Z" codes are divided into seven groups, addressing reasons for visits for examination and investigation: potential health hazards related to communicable diseases, circumstances related to reproduction, specific procedures, potential health hazards related to socioeconomic and psychosocial circumstances, other circumstances, potential health hazards related to family and personal history, and certain conditions influencing health status (Starfield, 1990). Many of the specific items in Chapter XXI were taken from previous editions of the ICD, mainly from the old "V" codes, but there are many new additions. The biggest problem, however, is that these types of codes, even when available, are rarely used by practitioners. An increasing focus on patients' problems, and incentives (perhaps even including financial incentives) for practitioners to record problems as well as diagnoses, may make these coding systems increasingly useful.

The World Organization of National Colleges and Academies of Family Practice has built on the work done over the past two decades by staff of the National Center for Health Statistics to develop a quite remarkable scheme for categorizing and coding problems; this too is little known and rarely used other than in some international research collaborations in office-based practice (Lamberts and Wood, 1987).

Thus, MEDTEP researchers already have much to draw on in their quest for better measures of risk to use as covariates in analyses. So far, these have been virtually ignored.

One serious remaining problem is the rule-out diagnosis, which Steinwachs (in this volume) mentions. A possible reason for the inability to handle rule-out diagnoses in the ICD is its genesis as a method for categorizing cause of death. There is no such thing as rule-out cause of death! In hospital discharge data, rule-out diagnoses may occur but are unlikely, since hospitalization generally entails resolution of diagnostic dilemmas. The problem is serious in ambulatory care and especially in primary care, since uncertainty is characteristic. About one in nine diagnoses recorded by physicians caring for children have a prefix, and about half of these are "rule-out" (Starfield, 1990). The true frequency is undoubtedly higher, since practitioners tend to assign a diagnosis because one is usually expected; uncertainty may be a tacit admission of inadequate diagnostic skills. With fee-for-service reimbursement, a diagnosis is usually required. The inability to capture tentative diagnoses is a major impediment to any research or evaluation of effectiveness. Perhaps a refocus on problems and a legitimation of their coding will reduce the magnitude of this serious limitation to research.

Roos and colleagues (in this volume) address the issue of record linkages, which will be an increasing challenge as the focus shifts to outpatient issues, to problems, and to the concept of clinical status beyond the diagnosis in MEDTEP research. These investigators have years of experience in working with record linkage and have perfected it to a fine art. They know how to maximize matching and minimize errors in overmatching. Their experience, however, is facilitated by a system of care in which there is only one payer. Their conceptual challenges come from the need to link data systems that are nonclinical—that is, case registries, monitoring systems, and home care files—rather than from difficulties in linking standard claims forms and hospital discharge data. Their paper says little about the possibility of linkages to clinical records, although linkages to a data base concerning immunizations are mentioned. A fruitful area for research would take advantage of the willingness of those practitioners who are interested in computer technology and in collaborative practice-based research to engage in experiments to determine the potential for linking clinical information with claims information, hospitalization data, and vital records.

The accuracy of claims data to reflect clinical information was a subject of interest in several of the papers in this volume. The problems with accepting diagnostic information obtained from claims forms are legion. As long as providers are paid for diagnoses, there will be diagnosis creep, just as there is diagnosis-related group creep in hospital diagnoses (Simborg, 1981). Overdiag-

nosis rather than underdiagnosis is likely to be characteristic, but this will be partly compensated for in volume by the underrepresentation of diagnoses that are poorly reimbursed.

Unfortunately, this problem may not be solved with better access to medical records because the same incentives that lead to the underrecording of diagnoses on claims forms leads to the underrecognition of problems in clinical care. Where providers are not paid to deal with psychosocial problems, they are unlikely to make a concerted effort to recognize their existence in patients. Overreporting of problems and the justification for reported diagnoses are problems amenable to study by comparison of medical records and claims. As medico-legal documents, medical records are unlikely to have systematic errors concerning the provider's assessment of the patient's diagnoses. At least one medical effectiveness and patient outcomes project in the AHCPR portfolio addresses the concordance of claims and records. This is a study of medical practice variations in outpatient care, first controlling for differences in ambulatory case mix through application of the ACG system. As part of the study, which is attempting to determine the relationship between variations in resource use and quality of care, research collaborators in the State health department have assigned peer review organization (PRO) reviewers to outpatient settings to assess various aspects of the quality of care for several conditions: hypertension and diabetes in adults, asthma in adults and children, otitis media in children, and well-person care in adults and 2-year-old children. Information on diagnoses and use of laboratory services abstracted from the medical record will be compared with information on claims forms for reimbursement. The results of this study should provide MEDTEP researchers with at least an estimate of the accuracy of information from data sets based on billable claims.

Where Do We Go from Here?

The success of medical effectiveness research provides new insights concerning challenges for the future. Some of these challenges will be easy to meet, and some will be more difficult. The easy ones will involve continuing in the directions already laid out but will have increasingly more parallel paths involving increasing numbers of conditions. Many more problems lend themselves to MEDTEP, as witnessed by the continuing research applications from the field. The process of guideline setting—a related but separate undertaking of AHCPR and many other professional organizations—will itself generate new ideas for research.

As the field widens beyond its initial focus on Medicare and the elderly, new types of subjects will arise. Pediatricians, obstetricians, otolaryngologists, anesthesiologists, and other specialists will discover that they, too, have a mission, and there will be even more applications on specific topics. As the field widens to include these specialties, the importance of outpatient care and episodes of care that include presenting problems as well as diagnoses will become increasingly evident.

The importance of primary care within the health services system will become better recognized as the impetus of cost containment in managed care and gatekeeper systems converts to a focus on quality and effectiveness of care. This is what happened during the 1980s; the early emphasis on costs and the development of mechanisms to reduce them led to a situation where quality of care had to become an equal concern, and the rest is recent history.

Ambulatory care, both primary and secondary, will be found to have variations in practice patterns at least as striking as those across hospitals. These variations, combined with the need for financial accountability, will heighten the recognition of the need for data systems, the application of already existing systems to code problems, and the development of new systems to categorize codes so that they are useful for medical effectiveness research.

The organization and financing of this expanded portfolio are likely to be problematic. An enlarged constituency, which will accrue as a result of the increasing involvement of clinical medicine in effectiveness research, may help to garner increased resources for AHCPR. Added efficiency in organization of research efforts will come about through increased collaboration between the Agency and NIH. A need for such collaboration is perceived on both sides. Effectiveness research depends heavily on efficacy research of the type generated and supported by NIH. Efficacy studies are conducted under controlled conditions and often in very limited settings. The budget for such research tends to be high—much higher than could currently be envisaged by the Agency. Several of the institutes within NIH, with much larger pools of resources, have been venturing into effectiveness research and finding that their investigators have neither the knowledge nor the skills to undertake outcome assessments. Their studies are, like clinical trials, generally limited to highly selected institutions. Their findings require real-world testing of the sort that implies more population-based data. Thus, there is a natural synergism between what NIH does and what health services research does. Informal linkages between some of the Institutes—most notably the National Institute on Aging—have existed for a long time and are represented by joint review of proposals and shared funding. Formal linkages are now in

order to enhance the quality and generalizability of research directed at improving outcomes. The time for more formal linkages may be upon us.

It may also be time to marry medical effectiveness research to the broader field of health services research. Returning to the figure, we see that the concerns of medical effectiveness are part of a larger picture, encompassing not only the processes but also the structures of health systems. There is much more explanatory power to medical practice variations when organizational and financial factors are included. Health services researchers such as Epstein, Stern, and Weissman (1990) have contributed to our knowledge about variations resulting from access barriers and characteristics of financing. Older and now classical studies showed how the characteristics of hospitals are associated with important outcome differences, and the early studies of the quality of care in ambulatory settings documented the importance of prior training of physicians, the volume of their work, and their interrelationship within organizations. So far, MEDTEP projects have focused largely on the biomedical aspects for explanatory power and less on the characteristics of the setting. This is a logical area for expansion.

MEDTEP studies have often been limited to one age group and even to one gender. National research efforts are now beginning to redress the prior imbalance between research on men and research on woman. NIH has explicitly indicated its intent to focus more heavily on research concerning women's issues. Now federally funded researchers must signify that they have included minorities and women in research efforts when the research problem is relevant to these population groups.

Two other major population groups warrant similar attention: children and the socioeconomically disadvantaged. Many research topics are relevant to children, yet children are excluded. Often, this exclusion is not even recognized, as a perusal of the medical care literature will demonstrate; it is not uncommon for entire articles to report on a study and its results without even mentioning that only adults were included. There is little history of collaboration between pediatricians and internists, and there have only been a few instances of collaboration between internists and family physicians and between pediatricians and family physicians. The great success of MEDTEP in generating collaborations should augur well for future collaborations between child-oriented and adult-oriented specialists. Practice-based collaborative research is a logical mechanism for developing such collaborations. It does much better in avoiding the academic turf issues that seem to arise so frequently when research is generated from academic institutions. So far, most collaborative practice-based research has focused on clinical issues, a likely departure point given the traditional interests and training of practitioners in clinical issues only; but as time moves on and the imperatives become clear, these networks too will move into health services research.

Thus, the imperatives for the thinkers and guiders in MEDTEP are to foster linkages with NIH, encourage the development of ambulatory data sets and research in ambulatory settings, foster a focus on problems as well as on diagnoses, and give priority to research that has the greatest promise of representativeness.

It is very exciting to be party to the birth and early development of MEDTEP and to be able to review with great admiration the work done so far. The papers in this volume are a fit tribute to the field and an excellent stimulus to the next phase of challenges.

References

Agency for Health Care Policy and Research. (1990). *AHCPR program note: Medical treatment effectiveness research* (AHCPR Pub. No. 90-0059). Rockville, MD: Author.

Averill, R., Goldfield, N., McQuire, T., and others. (1991). *Design and evaluation of a prospective payment system for ambulatory care. Final report.* (Cooperative Agreement No 17-C-99369/1-02). Baltimore: Health Care Financing Administration.

Epstein, A., Stern, R., and Weissman, J. (1990). Do the poor cost more? A multi-hospital study of patients' socioeconomic status and use of hospital resources. *New England Journal of Medicine*, 322, 1122-1128.

Kaplan, S., Greenfield, S., and Ware, J.E., Jr. (1989). Impact of the doctor-patient relationship on the outcomes of chronic disease. In Stewart, M. and Roter, D. (Eds.), *Communicating with medical patients* (Interpersonal Communications No. 9). Newbury Park, CA: Sage Publications.

Lamberts, H. and Wood, M. (Eds.). (1987). *ICPC. International Classification of Primary Care.* World Health Organization of National Colleges, Academics, and Academic Associations of General Practitioners/Family Physicians. Oxford: Oxford University Press.

Simborg, D. (1981). DRG creep: A new hospital-acquired disease. *New England Journal of Medicine*, 304, 1602-1604.

Starfield, B. (1990). *Does Chapter XXI meet the needs for health care information in the 1990s?* Paper prepared for the CPHA symposium on ICD-10: Will it refocus health care information from death to health status? Chicago.

Starfield, B. (1992). *Primary care: Concept, evaluation, and policy.* New York: Oxford University Press.

Starfield, B. (In press). Child and adolescent health status measures. *Center for the Future of Children*, 2(2).

Starfield, B., Katz, H., Gabriel, A., and others. (1984). Morbidity in childhood: A longitudinal view. *New England Journal of Medicine*, 310, 824-829.

Starfield, B., Steinwachs, D., Morris, I., and others. (1979). Patient-provider agreement about problems. Influence on outcome of care. *Journal of the American Medical Association*, 242, 344-346.

Starfield, B., Wray, C., Hess, K., and others. (1981). The influence of patient-practitioner agreement on outcome of care. *American Journal of Public Health*, 71, 127-132.

Tenan, P., Fillmore, H., Caress, B., and others. (1988). PACs classifying ambulatory care patients and services for clinical and financial management. *Journal of Ambulatory Care Management*, 11, 36-53.

50272-101

REPORT DOCUMENTATION PAGE	1. REPORT NO. AHCPR92-0056	2.	3. Recipient's Accession No.

4. Title and Subtitle
Medical Effectiveness Research Data Methods

5. Report Date
July 1992

6.

7. Author(s) Mary L. Grady and Harvey A. Schwartz, Eds.

8. Performing Organization Rept. No.

9. Performing Organization Name and Address
DHHS, PHS, Agency for Health Care Policy and Research (AHCPR)
Information and Publications Division, 2101 E. Jefferson St.,
Rockville, MD 20852 Tel: 301-227-8364

10. Project/Task/Work Unit No.

11. Contract(C) or Grant(G) No.
(C) n/a
(G)

12. Sponsoring Organization Name and Address
Same as above

13. Type of Report & Period Covered
In-House

14.

15. Supplementary Notes
Use of funds for printing this report has been approved by the Director of the Office of Management and Budget through September 1992. AHCPR PUb. No. 92-0056

16. Abstract (Limit: 200 words)

Improvements in the quality, effectiveness, and cost-effectiveness of health care are the goal of medical treatment effectiveness/patient outcomes research. Outcomes research requires large amounts of data because it studies health care in the general population. Often, data obtained from different sources--such as, Medicare, hospitals, third-party payers--need to be combined but are incompatible. This volume presents innovative techniques that have been developed and tested by leaders in the field for using large data bases to carry out medical effectiveness research.

17. Document Analysis a. Descriptors
Publication of this document by AHCPR does not necessarily represent approval or official endorsement by the Agency for Health Care Policy and Research or the U.S. Department of Health and Human Services.

b. Identifiers/Open-Ended Terms
Medical effectiveness research; MEDTEP; patient outcomes; record linkage; data; episodes of care; survival analysis

c. COSATI Field/Group

18. Availability Statement: Releasable to the public. Available from the National Technical Information Service, Springfield, VA Tel: 703/487-4650	19. Security Class (This Report) Unclassified	21. No. of Pages 190
	20. Security Class (This Page) Unclassified	22. Price

(See ANSI-Z39.18)

OPTIONAL FORM 272 (4-77)
(Formerly NTIS-35)
Department of Commerce

CMS LIBRARY
3 8095 00000519 5

DEPARTMENT OF HEALTH & HUMAN SERVICES

Public Health Service
Agency for Health Care Policy and Research
Executive Office Center
2101 East Jefferson Street, Suite 501
Rockville, MD 20852

AHCPR92-0056